NLP Question Bank

Q1. Brief about WordNet structure with example.

Ans:

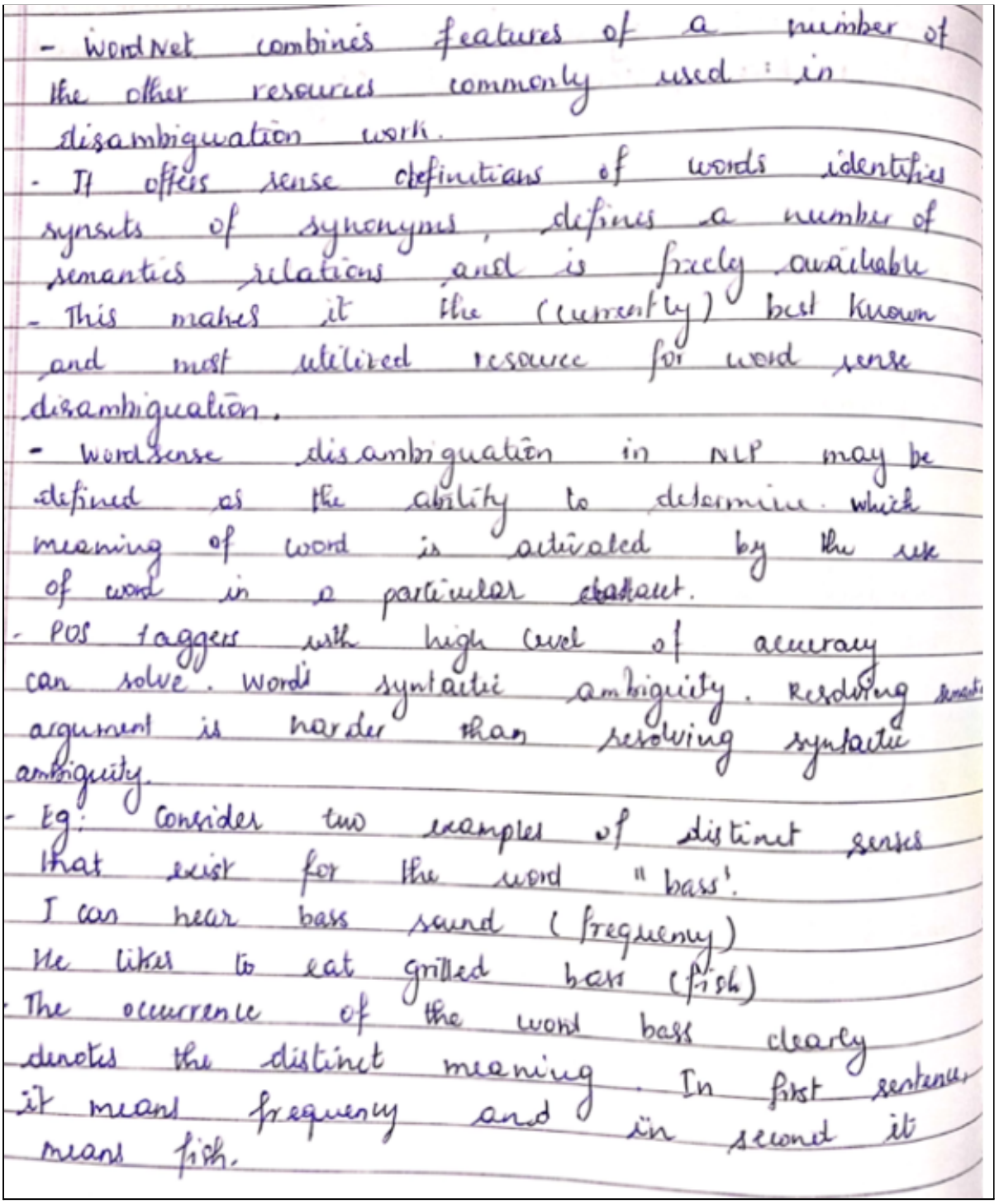
- Wordnet is a big collection of words from the English language that are related to each other and are grouped in some way.
- It is also called a lexical database.
- In other words, WordNet is a database of English words that are connected together by their semantic relationships.
- It is like a superset dictionary with a graph structure.
- WordNet groups nouns, verbs, adjectives, etc. which are similar and the groups are called synsets or synonyms.
- In a wordnet a group of synsets may belong to some other synset.
- For example, the synsets "Stones" and "cement" belong to the synset "Building Materials" or the synset "Stones" also belongs to another synset called "stonework".
- In the example given, stones and cement are called hyponyms of synset building materials and also the synsets building materials and stonework are called synonyms.
- Every member of a synset denotes the same concept but not all synset members are interchangeable in context.
- The membership of words in multiple synsets or concepts mirrors polysemy or multiplicity of meaning.
- There are three principles the synset construction process must adhere to:
    a. Minimality:
        - This principle determines on capturing those minimal set of the words in the synset which especially identifies the concept.
        - For example, (family, house) uniquely identifies a concept e.g, "she is from the house of the Classical Singers of Hyderabad".
    b. Coverage:
        - This principle's main aim is completion of the synset, that is capturing all those words that represent the concept expressed by the synset.In the synset, the words should be ordered according to their frequency in the collection.

c. Replaceability:

- Replaceability dictates that the most common words, in the synset, that is words towards the beginning of the synset should be able to replace one another.

Q2. Write a short note on word disambiguation.

Ans:


- WordNet combines features of a number of the other resources commonly used in disambiguation work.
- It offers sense definitions of words identifies synsets of synonyms, defines a number of semantics relations and is freely available
- This makes it the (currently) best known and most utilized resource for word sense disambiguation.
- Word sense disambiguation in NLP may be defined as the ability to determine which meaning of word is activated by the use of word in a particular context.
- POS taggers with high level of accuracy can solve Word's syntactic ambiguity. Resolving semantic argument is harder than resolving syntactic ambiguity.
- Eg: Consider two examples of distinct senses that exist for the word "bass".
I can hear bass sound (frequency)
He likes to eat grilled bass (fish)
The occurrence of the word bass clearly denotes the distinct meaning. In first sentence it means frequency and in second it means fish.

- There are four conventional approaches to WSD:

1. Dictionary and knowledge-based methods: These rely primarily on dictionaries, thesaurus , and lexical knowledge bases, without using any corpus evidence.
2. Supervised methods: These make use of sense-annotated corpora to train from.
3. Semi-supervised or minimally-supervised methods: These make use of a secondary source of knowledge such as a small annotated corpus as seed data in a bootstrapping process, or a word-aligned bilingual corpus.
4. Unsupervised methods: These eschew (almost) completely external information and work directly from raw unannotated corpora. These methods are also known under the name of word sense discrimination.

Q3. Explain discourse - reference resolution.

Ans:

- The most difficult problem of AI is to process the natural language by computers or in other words *natural language processing* is the most difficult problem of artificial intelligence.
- If we talk about the major problems in NLP, then one of the major problems in NLP is discourse processing – building theories and models of how utterances stick together to form coherent discourse.
- Actually, the language always consists of collocated, structured and coherent groups of sentences rather than isolated and unrelated sentences like movies. These coherent groups of sentences are referred to as discourse.
- Coherence and discourse structure are interconnected in many ways.
- Coherence is used to evaluate the output quality of a natural language generation system.
- Interpretation of the sentences from any discourse is another important task and to achieve this we need to know who or what entity is being talked about. Here, interpretation reference is the key element.
- Reference may be defined as the linguistic expression to denote an entity or individual.

- For example, in the passage, Ram, the manager of ABC bank, saw his friend Shyam at a shop. He went to meet him, the linguistic expressions like Ram, His, He are referenced.
- Reference resolution may be defined as the task of determining what entities are referred to by which linguistic expression.

Q4. Which algorithm is used for pronoun resolution in pragmatic analysis?

Ans:

- Hobbs' algorithm was one of the earliest approaches to pronoun resolution.
- The algorithm is mainly based on the syntactic parse tree of the sentences. It makes use of syntactic constraints when resolving pronouns.
- Hobbs' algorithm prefers entities that are within the same sentence, and entities that are closer pronoun in the same sentence.
- Depending on the position of the pronoun in the sentence, different entities in a sentence may become more relevant.
- When looking for antecedents in previous sentences, the antecedents that occur (or are realised) in the subject position are more salient, since a breadth-first left-to-tree search is performed starting at the root S node of the sentence.
- Depth of a node in the syntactic tree is thus a very important factor to determine discourse prominence.
  1. Start with target pronoun
  2. Climb parse tree to S root
  3. For each NP or S:
     a. Do breadth-first, left-to-right search of children
     b. Restricted to left of target
     c. For each NP, check agreement with target

Q5. How NLP is used for text categorization and text summarization?

Ans:

Text Categorization:

- Text classification also known as text tagging or text categorization is the process of categorising text into organised groups.

- By using Natural Language Processing (NLP), text classifiers can automatically analyse text and then assign a set of pre-defined tags or categories based on its content.
- Approaches:
- Text Classification can be achieved through three main approaches:

1. Rule-based approaches:
    - These approaches make use of handcrafted linguistic rules to classify text.
    - One way to group text is to create a list of words related to a certain column and then judge the text based on the occurrences of these words.
    - For example, words like "fur", "feathers", "claws", and "scales" could help a zoologist identify texts talking about animals online.
    - These approaches require a lot of domain knowledge to be extensive, take a lot of time to compile, and are difficult to scale.
2. Machine learning approaches:
    - We can use machine learning to train models on large sets of text data to predict categories of new text.
    - To train models, we need to transform text data into numerical data - this is known as feature extraction.
    - Important feature extraction techniques include bags of words and n-grams.
    - There are several useful machine learning algorithms we can use for text classification.
    - The most popular ones are Naive Bayes classifiers, Support vector machines, Deep learning algorithms
3. Hybrid approaches:
    - These approaches are a combination of the two algorithms above.
    - They make use of both rule-based and machine learning techniques to model a classifier that can be fine-tuned in certain scenarios.

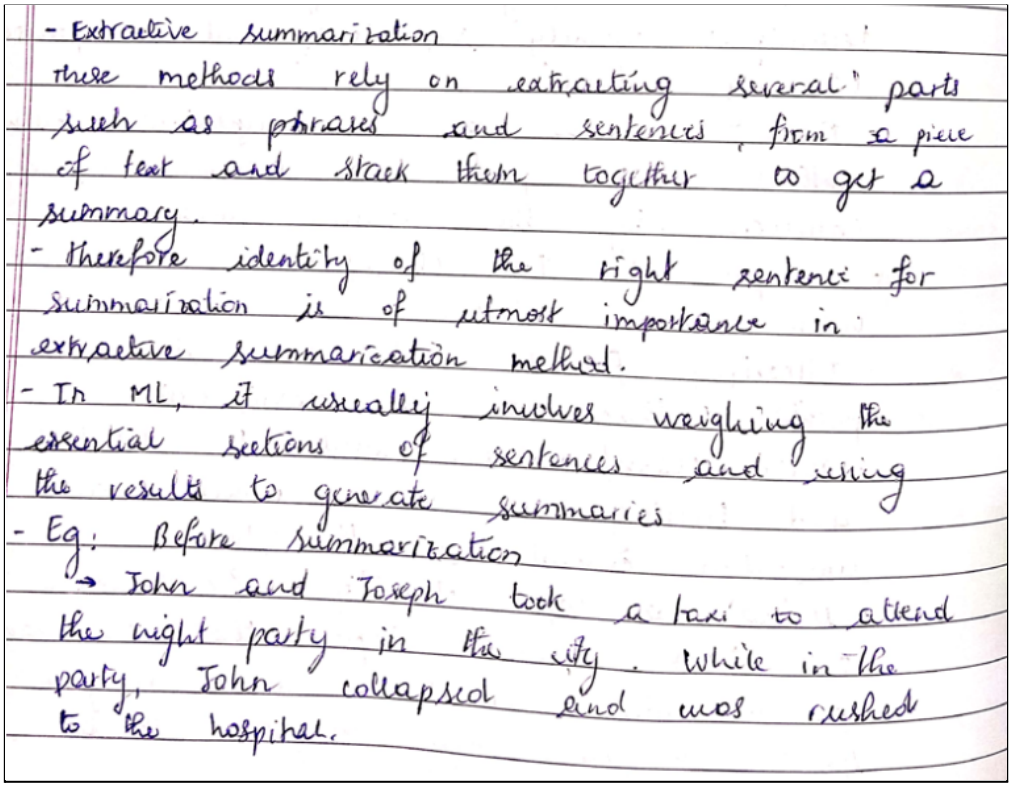
- Some of the most common examples and use cases for automatic text classification include the following:

- Sentiment Analysis: the process of understanding if a given text is talking positively or negatively about a given subject (e.g. for brand monitoring purposes).
- Topic Detection: the task of identifying the theme or topic of a piece of text (e.g. know if a product review is about Ease of Use. Customer Support, or Pricing when analysing customer feedback).
- Language Detection: the procedure of detecting the language of a given text (e.g. know if an incoming support ticket is written in English or Spanish for automatically routing tickets to the appropriate team).

Text Summarization:

- A summary is a reductive transformation of a source text into a summary text by extraction or generation.
- The goal of summarization is to produce a shorter version of a source text by preserving the meaning and the key contents of the original document.
- A well written summary can significantly reduce the amount of work needed to digest large amounts of text.
- There are two types summaries

1. Extractive summaries
2. Abstractive summaries



After summarization
→ John and Joseph attend party.
John rushed hospital.

Abstractive Summarization
- Here summary of the texts can be different from the original text, which is contrast to extraction based summarization where only existing sentences which are present are used.
- Advanced deep learning techniques are used to generate the new summary.
- Eg. Before Summarisation
→ John and Joseph took a taxi to attend the night party in the city. While in the the party, John collapsed and was rushed into the hospital
After summerization
→ John was hospitalized after attending the party.

Q6. Short Note on Information Retrieval

Ans:

- Information retrieval 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.
- An information retrieval system searches a collection of natural language documents with the goal of retrieving exactly the set of documents that matches a user's question.
- They have their origin in library systems.
- These systems assist users in finding the information they require but it does not attempt to deduce or generate answers.
- It tells about the existence and location of documents that might consist of the required information that is given to the user.

- The documents that satisfy the user's requirement are called relevant documents.
- If we have a perfect IR system, then it will retrieve only relevant documents.
- A user who needs information will have to formulate a request in the form of a query in natural language.
- After that, the IR system will return output by retrieving the relevant output, in the form of documents, about the required information.
- The step by step procedure of these systems are as follows:
    - Indexing the collection of documents.
    - Transforming the query in the same way as the document content is represented.
    - Comparing the description of each document with that of the query.
    - Listing the results in order of relevancy.
- Retrieval Systems consist of mainly two processes:
    - Index
    - Matching

Q7. Explain how maximum entropy and CRF is used in NLP.

Check Assignment

Q8. Describe Homonymy, Polysemy, Hyponymy and Antonymy with Example

Ans:

Homonomy:
- A partial relation that holds between two lexemes that have the same form but unrelated meaning.
- complicated by presency of two kinds of forms
- ex:
  1) bat (wooden stick) vs bat (flying mamal)
  2) bank (financial institution) vs bank (riverside).

Polysemy:
- A single lexemes with mulliple related senses
- Polysemy refers to words or phrases with different but related meanings. A word becomes polysemous if it can be used to express different meanings
- It can be sometimes difficult to determine whether a word is polysemous or not because the relations between the words can be vague and unclear
- Eg: She wore a simple duress (undecorrated dress)
  The sum was simple (easy).

Hyponymy.

- Hyponymy is the state or phenomenon that shows relationship between more general term (lexical representation) and the more specific instance of it)
- The concrete forms of set of words are called hyponyms
- eg: 1) car, bike, cycle is vehicle
  So we can say car is hypony of vehicle

  2) red, blue, green, is colour
  So we can say red is hyponym of colour

Antonymy.

- It is the state or phenomenon in which words have the sense relation which involve the opposite of meaning
- word pairs of antonym can be divided into several types:
  - → gradable antonyms
  - → complementary antonyms
  - → relational antonyms
- gradable antonyms are pair of words with opposite that lie on a continuous spectrum:
  - for eg: young ; old,
    good ; bad
- complementary pairs refers to the existence of pairs that the denial of one, implies the assertion of the other.
  for eg: dead ; alive male ; female