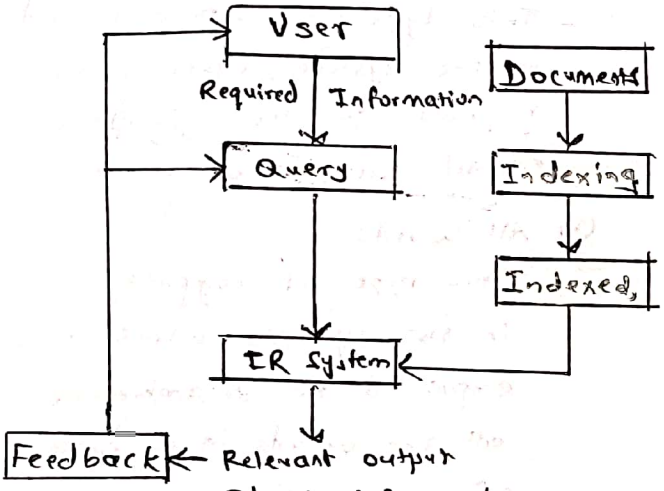
## Information Retrieval

- IR is defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information.
- The system helps users in finding the information they require but it does not explicitly provide answer to the question.
- It informs the existence and location of documents that might consist of the required info.
- Document that satisfies user's requirement are called relevant documents
- A perfect IR system will retrieve only relevant documents.
- Structure of IR:




- Classical Problem in IR system
  ⇒ Ad-hoc retrieval problem
- In Ad-hoc retrieval, the user must enter a query in natural language that describes the required information. Then the IR system will return the required documents related to desired information.
- For ex, Suppose we are searching something on the internet and it gives some exact pages that are relevant as per requirement but there can be some irrelevant pages too. This is due to adhoc retrieval problem.

- IR model
  - A model of IR predicts and explain what a user will find in relevance to the given query.
  - IR model consists of:
    1. A model for document
    2. A model for queries
    3. A matching function that compares queries to documents.
- Types of IR models
  1. Classical IR model
     - Simple and easy to implement
     - Based on mathematical knowledge that was easily recognized and understood.
     - Eg. Boolean, Vector, Probability
  2. Non-Classical IR model
     - Opposite of classical IR model.
     - Based on principles other than similarity, probability, boolean op".
     - Eg. Information logic model, Situation Theory model, Interaction model.
  3. Alternative IR model
     - It is an enhancement of classical IR model.
     - Eg, Cluster model, Fuzzy model, Latent semantic Indexing (LSI)

## Word Sense Disambiguation

- WSD is a well known problem in NLP
- WSD is used in identifying what the sense of word means in a sentence when the word has multiple meanings.
- When a single word has multiple meaning, then for the machine it is difficult to identify the correct meaning and to solve this challenging issue we can use the rule-based system or machine learning techniques.
- WSD is a natural classification problem: Given a word and its possible senses, as defined by a dictionary, classify an occurrence of the word in context in one or more of its sense classes.
- Example:
  "I saw her duck."

  Here, WordNet lists two senses for the word saw.
  1. saw - The action of having seen something. (seeing her duck).
  2. saw - The action of using a chainsaw. (cutting her duck)

- WSD Methods
  1. Dictionary and knowledge-based methods.
     - These methods rely on text data like dictionaries, thesaurus, etc.
     - It is based on the fact that words that are related to each other can be found in the definitions.
  2. Supervised methods
     - In this, sense-annotated corpora are used to train machine learning models.
     - Only problem is, such corpora are difficult to create.
  3. Semi-supervised methods
     - Due to lack of such corpora, most word sense disambiguation algorithms use semi-supervised method
     - The process starts with small amount of data, which is also called as seed data
  4. Unsupervised method
     - This is the greatest challenge to researchers and NLP professionals
     - A key assumption is that similar meanings and senses occur in a similar context. They are not dependent on manual efforts.

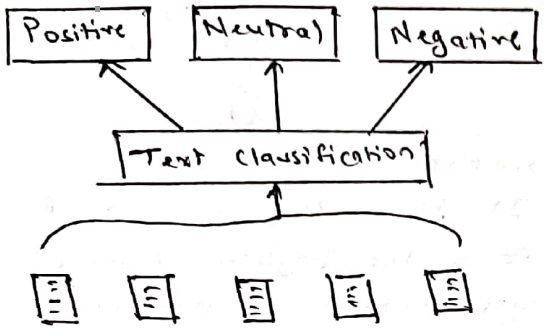
- WSD Evaluation
  - Evaluation of WSD requires two inputs:
  1. A Dictionary
     - First input for WSD evaluation,
     - It is used to specify the senses to be disambiguated,
  2. Test Corpus
     - Another input for WSD evaluation
     - It can be of two types:
       a. Lexical Sample
          - This type of corpora is used in the system, where it is required to disambiguate a small sample of words.
       b. All Words
          - This type of corpora is used in the system, where it is expected to disambiguate all the words in a piece of running text.

- Difficulties in WSD
  1. Difference between dictionaries and text corpus as different dictionaries have different meaning of words.
  2. Different application needs different algorithms.
  3. Words often have related meanings so sometimes word cannot be divided into discrete meanings

- Applications of WSD
  1. Machine Translation
  2. Text mining and information extraction
  3. Information retrieval
  4. Lexicography

## Text Categorization

- Text classification also known as text tagging or text categorization.
- It is a process of categorizing text into organized groups.
- By using NLP, text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.
- e




Book Review

- Text categorization approaches

① Rule-based approach
- It uses handwritten rules to classify text.

② Machine Learning approaches
- It is used to train models on large text dataset to predict category of new text.
- ML algorithms used for text classification are:
  - ⓐ Naive Bayes Classifier
  - ⓑ Support Vector Machines
  - ⓒ Deep Learning Algorithms

③ Hybrid approaches
- It is a combination of above two approaches.
- They make use of both rule-based and ML techniques to model a classifier that can be fine tuned in certain scenarios.

- Text Categorization Examples.

① Sentiment Analysis
- The process of understanding if a given text is talking positively or negatively about a given subject
- Eg, social media monitoring

② Topic Detection
- The task of identifying the theme or topic of a piece of text.
- Eg, Know if a product review is about ease of use or pricing when analyzing feedback

③ Language Detection
- Procedure of detecting a language of a given text.

## Text Summarization

- A summary is a reductive transformation of a source text into a summary text by extraction or generation.
- Goal of summarization is to produce a shorter version of source text by preserving the meaning and key content of original document.
- Types of text summarization

① Extractive Summarization
- Extractive summaries are created by reusing portions (words, sentences) of the input text document.
- The system extracts text from the entire collection, without modifying text document.
- Most of the summarization research today is on extractive summarization.

② Abstractive Summarization
- It requires deep understanding and reasoning over the text.
- It provides own summary over input text without using same word or sentence from the input text.
- Determines the actual and short meanings of each element, such as words, sentences, paragraphs, etc.
- It is more efficient than extraction
- Abstractive summarization algorithms are complex to build.

## Verb Phrase (VP)

- English VPs consist of a verb (the head) along with 0 or more following constituents.:

VP → verb — disappear

VP → Verb NP — prefer a morning flight

VP → Verb NP PP — Leave boston in the morning

VP → Verb PP — Leaving on Thursday

Note:

NP → Noun phrase

PP → prepositional phrase

## Hyponymy

- One sense is a hyponym of another if the first sense is more specific, denoting a subclass of the other.
- Eg, Car is hyponym of vehicle

## Polysemy

- A single lexeme with multiple related meanings.
- Eg, Date (fruit) vs
  Date (a particular day)

## Homonymy

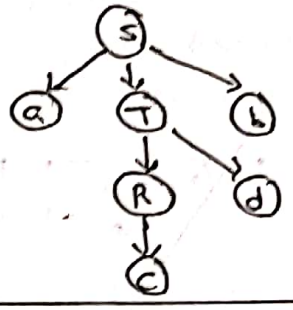
- Lexemes that share a form, but have unrelated, distinct meanings.
- Eg. Bat (Wooden stick like thing)
  vs
  Bat (Flying scary mammal)

## Antonymy

- Senses that are opposites with respect to one feature of their meaning.
- Eg. Dark / Light
  Short / Long
  Hot / Cold

## Top-down Parser

- Also known as recursive parsing or predictive parsing.
- In Top-down parsing, the parsi start from the start symbol and transform it into input symbol
- Parse tree representation of inpu string 'abcd' is:




## WordNet

- It is a big collection of words from the English language that are related to each other and are grouped in some way.
- Also called as lexical database.
- WordNet is a database of english words that are connected together by their semantic relationships.
- It is like superset dictionary with graph structure.
- WordNet groups nouns, verbs, adjectives, etc. which are similar and groups are called synonyms. Or synset.
- In WordNet, group of synsets may belong to other synset.
- Eg. Synsets "stones" and "cement" belong to synset "Building material". Synset "stone" also belong to synset "stonework".
- WordNet is open source and available for download

Discourse - Reference Resolution

- The most difficult problem of AI is to process the natural language by computers.
- NLP is the most difficult problem of AI.
- If we talk about major problems in NLP, then one of the major problems 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 as discourse.
- Coherence and discourse structures 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 imp 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.
- Reference resolution is defined as the task of determining what entities are referred to by which linguistic expression.
- For example,

  Ram, the manager of ABC bank saw his friend Shyam at a shop. He went to meet him.

  Here, the linguistic expressions like Ram, His, He are referenced.

- Hobb's algorithm is used for pronoun resolution in pragmatic analysis.
- Hobb's algorithm was one of the earliest approaches to pronoun resolution
- Algorithm is based on syntactic parse tree of the sentences. It makes use of syntactic constraints when resolving pronouns.
- Hobb's 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 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 the node in the syntactic tree is very important factor to determine discourse prominence.

  ① Starting with target pronoun.
  ② Climb parse tree to S root.
  ③ For each NP or S:
    ⓐ Do breadth first search, left-to-right search for children.
    ⓑ Restricted to left of target.
    ⓒ For each NP, check agreement with target.

## Maximum Entropy

- It is a framework for integrating information from many heterogenous information source for classification
- The term maximum entropy refers to an optimization framework in which the goal is to find the probability model that maximizes entropy over the set of models that are consistent with the observed evidence.
- It is used to predict observations from training data. This does not uniquely identify the model but chooses the model which has the most uniform distribution. i.e the model with the maximum entropy

## Conditional Random Fields

- CRF is conditional probabilistic model for sequence labelling just as structured perception is built on the perception classifier, conditional random fields are built on the logistic regression classifier
- CRFs are a probabilistic framework for labelling and segmenting sequential data based on the conditional approach.
- CRF is a form of undirected graphical model that defines a single log linear distribution over label sequences given a particular observation sequence