Module 1
Introduction to NLP

NLP
- It stands for Natural Language Processing
- It is a part of Computer science, Human Language and Artificial Intelligence.
- It is a field of study that focuses on the interactions between human language and computers.
- It is the technology that is used by machines to understand, analyze, manipulate and interpret human language.
- Types of NLP
  1. Natural language recognition
  2. Natural language generation
- Need of NLP
  1. Better understanding of human communication
- Example : Facebook uses NLP to track trending topics & popular hashtags

Applications of NLP
1. Machine Translation
2. Speech Recognition
3. Speech synthesis
4. Information Retrieval
5. Information Extraction
6. Question Answering
7. Text summarization
8. sentiment Analysis

Challenges of NLP
1. Contextual words and phrases and homonyms
- The same words and phrases can have diverse meanings according the context of a sentence and many words have the exact same pronunciation but completely different meanings.
- For Example:
a. I ran to the store because we ran out of milk.
b. Can I run something past you quickly?
- In the above two sentences the meaning of the run is different according to the content.
- Homonyms means the pronunciation of two or more words is same but have different meanings.
- Examples, Their and There, Right and Write

2. Synonyms
- It can cause issues like contextual understanding since we use many different words to express the identical idea.
- Additionally, some of these words may convey exactly the same meaning, while some may be levels of complexity
- For ex; small, little, tiny, minute

3. Irony and sarcasm
- It creates problem for machine learning models since they usually use words and phrases that, strictly by definition, may be positive or negative but truly mean the opposite
- Models can be trained with certain indications that frequently accompany ironic or sarcastic phrases.
- Example; like yeah right, whatever, etc.

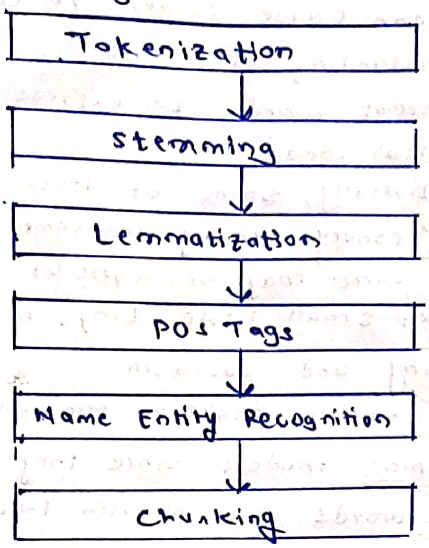
4. Ambiguity
- It refers to sentences and phrases that potentially have two or more possible interpretations.
- There is lexical, syntactic and semantic ambiguity

5. Errors in text or speech
- Misspelled or misused words can create problem for text analysis.
- Autocorrect and grammar correction applications can handle common mistakes but do not at all times understand the writer's intention.

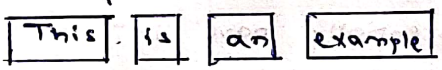
6. Idioms and slang
- Informal phrases, expressions, idioms and culture-specific lingo present a number of problems for NLP.

7. Domain specific language
- Different business and industries use very different language.
- An NLP processing model needed for healthcare would be very different than one used to process legal documents.

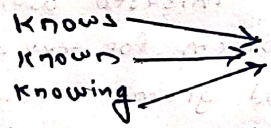
Text Processing steps in NLP



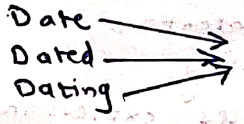
① Tokenization
- Tokenization is the process of exchagging sensitive data for non-sensitive data called 'Tokens' that can be used in a database.
- It is a process of breaking sentence into small tokens.
- Example: This is an example.



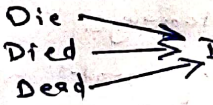
② Stemming
- It is a process of normalizing words in their root form.
- Example:



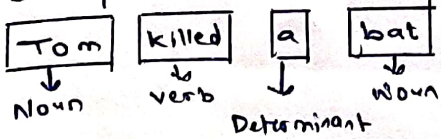
- Stemming is not accurate and efficient
- Example:



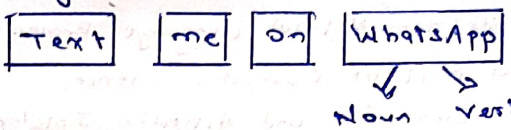
③ Lemmatization
- It is an extension of stemming
- Output of lemmatization is accurate
- It combines similar words to generate output (Lemma).
- Example:



④ POS Tags (Part of speech Tags)
- It tags all the tokens with POS Tags.
- Example



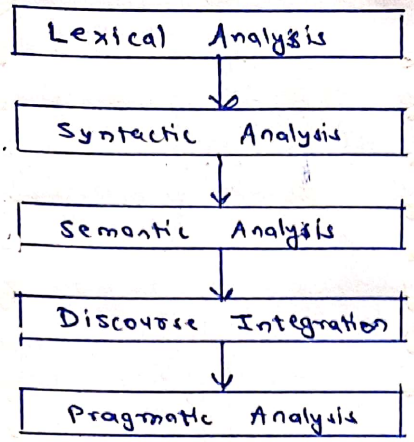
Disadvantage:
- There could be multiple POS Tags to single token.



⑤ Name Entity Recognition
- It overcomes drawbacks of POS Tags
- It involves identification of key information in the text and classify it into predefined categories.
- Some of the categories includes person, organization, place/location, etc.

⑥ Chunking
- It groups individual tokens and forms a chunk.
- It helps to get insightful and meaningful information from text.

Stages of NLP



- There are 5 stages in NLP.

① Lexical Analysis
- It is the first stage in NLP.
- It is also known as morphological analysis.
- In this stage, structure of words is identified and analyzed
- It divides whole text into paragraphs, sentences and words.

② Syntactic Analysis
- It is also known as Parsing.
- It is used to check grammar, word arrangements and shows relationship among the words.
- Example : Agra goes to Rutuja
  In the real world, Agra goes to Rutuja does not make any sense, so this sentence is rejected by syntactic analyzer.

③ Semantic Analysis
- It mainly focuses on literal meaning of words.
- It draws exact dictionary meaning from text.
- Text is checked for meaningfulness
- It is done by mapping syntactic structure.

④ Discourse Integration
- The meaning of any sentence depends upon the meaning of the sentence just before it.
- It brings meaning for immediate following sentence
- In the text, "Archit is a bright student. He spends most of the time in the library." Here, discourse assigns "He" to refer to "Archit".

⑤ Pragmatic Analysis
- It is the fifth and last stage of NLP.
- In this stage, what was said is re-interpreted on what is truly meant.
- It contains deriving those aspects of language which necessitate real world knowledge.
- For example, "Open the book" is interpreted as a request instead of an order.

Module 2
Word level analysis

N-gram language model

- A language model in NLP is a probabilistic statistical model that determines the probability of a given sequence of words occuring in a sentence based on the previous words.
- It helps to predict which word is more likely to appear next in the sentence.

N-gram

- It is defined as the contiguous sequence of n items from a given sample of text or speech.
- The items can be letters, words, etc.
- The N-grams are typically collected from a text or speech corpus.
- Examples of N-gram
  1. Unigram
     ("This", "is", "an", "example")
  2. Bigram
     ("This is", "is an", "an example")
- N-gram is used for error detection and spelling correction.
- N-gram is used without a dictionary, this way employs to find in which position the error occured in an incorrect word.
- If there is any way to change the incorrect word so that it contains only correct N-grams, then N-gram is used for correction.

Regular Expression

- Also called as Regex.
- It is a language used for specifying text search string.
- It is a powerful way to find and replace string that take a defined format. For ex; regex are used to parse email addresses, url's, dates, log files, programming script, etc.
- Regex is a useful tool to design language compiler as well as they are used in natural language processing for tokenization, morphological analysis etc.
- The simple type of regular expression contains a single symbol.
- It also specifies sequence of characters.

1. Brackets:
   - Characters are grouped by putting them between square brackets.
   - Example,
     /[abcd]/ will match any of a, b, c and d.
     /[0123456789]/ specifies any single digit.
2. Range:
   - Sometimes regular expression lead to cumbersome notation
   - Example,
     /[abcdefghijklmnopqrstuvwxyz]/
     It specifies any lowercase character.
   - In such cases a dash is used to specify a range.
   - Example,
     /[3-6]/ - Specifies any one of the digits 3, 4, 5, 6.
     /[c-f]/ - specifies any one of the letter c, d, e, f.
3. Caret ^:
   - It is used at the beginning of a regular expression to specify what a single character cannot be.
   - Example,
     /[^x]/ - matches any single character except x.
     /[^A-Z]/ - Not an uppercase letter.
     /[^Tt]/ - Neither "T" nor "t".
   - Regex are case sensitive.
4. * or +:
   - The use of * or + allows you to add 1 or more of a preceding character.
   - Example,
     paat → paa, paaaa, paaaaa.
5. ? - The question mark:
   - Letters optionality of the previous expression
   - Example,
     /woodchucks?/ - woodchuck or woodchucks.

Inflectional Morphology

1. It is a morphological process that adopts existing words so that they function effectively in sentences without changing pos of base morpheme.
2. They close off the word. eg. Plays.
3. They cannot be found in dictionaries.
4. Inflection is relevant to syntax.
5. It is obligatory
6. It expresses the same concept as base
7. It is semantically regular
8. It is expressed closed to the root.
9. Meanings are less relevant to the meaning of the base.
10. Meanings are relatively abstract
11. It can be suffix or infix but not prefix.
12. Example,

cat (Noun) + s = cats (Noun)

Derivational Morphology

1. It is concerned with the way morphemes are connected to existing lexical forms as affixes.
2. They never close off the word. eg. playful
3. They can be found in dictionaries.
4. Derivation is irrelevant to syntax.
5. It is optional.
6. It expresses a new concept.
7. It is semantically irregular.
8. It is expressed at the node of words.
9. Meanings are relevant to the meaning of the base.
10. Meanings are relatively concrete
11. It can be prefix or suffix.
12. Example,

danger (Noun) + ous = dangerous (Adjective)

Stemming

1. Stemming is faster as it chops words without knowing the context of the sentence.
2. It is a rule based approach.
3. Accuracy is less
4. When we convert any word to its root form then stemming may create the non-existent meaning of word
5. Stemming is preferred when meaning of the word is not important for analysis.
   For example, Spam detection
6. Example,
   "studies" ⟶ "studi"

Lemmatization

1. Lemmatization is slower as compared to stemming but it knows the context of the word before proceeding
2. It is a dictionary based approach
3. Accuracy is more
4. Lemmatization always give the dictionary meaning of word while converting into root form.
5. Lemmatization would be recommended when the meaning of word is important for analysis
   For example, Question Answering
6. Example,
   "studies" → "study"

Porter's Stemming Algorithm

- It is one of the most popular stemming methods proposed in 1980.
- It is based on the idea that the suffixes in the English language are made up of a combination of smaller and simpler suffixes.
- This stemmer is known for its speed and simplicity.
- The main applications of porter stemmer include data mining and information retrieval.
- However, its applications are only limited to English words.
- Also, the group of stems is mapped onto the same stem and the output stem is not necessarily a meaningful word.
- The algorithms are fairly lengthy in nature and are known to be the oldest stemmer.
- Example,
  EED → EE means "If the word has at least one vowel and consonant + EED ending, change the ending to EE."
  as "agreed" becomes "agree".

Advantage

- It produces the best output as compared to other stemmers and it has less error rate.

Limitations

- Morphological variants produced are not always real words.

Finite state Automata (FSA)

- An automata having finite number of states is called a Finite Automata or Finite state Automata (FSA).
- FSA is used to recognize patterns.
- It takes the string of symbol as input and changes its state accordingly.
- When the required symbol is found then the transition happens.
- When transition takes place, the automata can either move to the succeeding state or stay in the current state.
- There are two states in FSA.
  1. Accept
  2. Reject
- When the input state is processed successfully and the automata reached its final state, then it will accept.
- Mathematically, an automata can be represented by 5-tuples
  $(Q, \Sigma, \delta, q_0, F)$

  where,

  Q = Finite set of states

  Σ = Finite set of symbols

  δ = Transition Function

  q_0 = Initial state

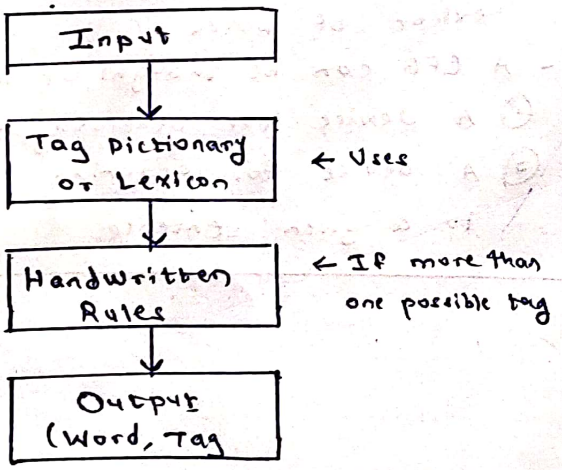
  F = Final state

Module 3
Syntax Analysis

Part of Speech Tagging
- It is a process of marking up a word in a text as corresponding to a particular part of speech.
- It a process of converting a sentence to forms. (words, POS Tags)
- POS are divided into two categories:
① Open class → Noun, verb, adjective, adverb.
② Close class → preposition, determiner, conjuction, pronouns, particples.
- POS Tagging is majorly used in
① Text to speech
② Parser
③ Search bar
- Biggest challenge in POS Tagging is Ambiguity

Methods for POS Tagging
- Most tagging algorithms fall into one of two classes:
① Rule-based taggers
② Stochastic taggers

Rule-Based POS Tagging



Properties of Rule-based POS Tagging
① Taggers are knowledge driven.
② Rules are built manually.
③ Information is coded in the form of rules.
④ We have approximately 1000 rules.
⑤ Smoothing and language modelling is defined explicitly.

Stochastic POS Tagging
- The model that includes frequency or probability (statistics) can be called stochastic.
- There are two approaches in stochastic POS Tagging.

① Word Frequency Approach
- The tag encountered most frequently in the training set is the one assigned to the ambiguous instance of the word

② Tag sequence probabilities
- The best tag for a given word is determined by the probability that it occurs with the n previous tags.
- It is also called as n-gram approach.

Properties of stochastic POS Tagging
① The POS Tagging is based on the probability of the tag occuring.
② It requires training corpus.
③ There would be no probability for the words that do not exist in the corpus.
④ It uses different testing corpus other than training corpus.
⑤ It is the simplest POS tagging because it chooses most frequent tags associated with a word in training corpus.

Hidden Markov Models (HMM)

- HMMs are sequence models.
- HMMs are a statistical markov model in which the system being modelled is assumed to be a markov process with an unobservable (hidden) states.
- They are designed to model the joint distribution $P(H, O)$, where H is the hidden state and O is the observed state.
- For example, in the context of POS tagging, the objective would be to build a HMM to model P(word, tag) and compute the label probabilities given observations using Bayes' rule.

$$P(H|O) = \frac{P(H|O)\,P(H)}{P(O)}$$

- HMM graph consists of hidden space and observed space where the hidden space consists of the labels and the observed space is the input.
- These spaces are connected via transition matrices $\{T, A\}$ to represent the probability of transitioning from one state to another following their connections.
- Each connection represents a distribution over possible options; given our tags this results in a large search space of the probability of all words given the tag.
- The main idea behind HMM is that of making observations and travelling along connections based on a probability distribution.
- In the context of sequence tagging, there exists a changing observed state (the tag) which changes as our hidden state (tokens in the source text) also changes.

Context Free Grammar (CFG)

- Also called as Phrase-structure grammar
- CFG is equivalent to Backus-Naur Form (BNF).
- CFG's are powerful enough to describe most of the structure in natural languages.
- CFG's are restricted enough so that efficient parsers can be built.
- CFG is a notation for describing languages and a superset of Regular Grammar.
- CFG is a formal grammar which is used to generate all possible strings in a given formal language.
- CFG can be described by 4 tuples.

  $G = (V, T, P, S)$

  where,

  G = grammar

  V = Finite set of non-terminal symbols.

  P = production rules

  S = start symbol

- A CFG consists of a set of rules or production, each expressing the ways the symbols of the language can be grouped together and a lexicon of words.
- A CFG can be thought of in 2 ways
  1. A device for generating sentences.
  2. A device for assigning a structure to a given sentence.

Open Class

① It does not have relatively fixed membership.
② It is also called as content words.
③ It includes noun, verb, adjective, adverb.
④ Here, words are infinite
⑤ They are mutually inclusive
⑥ Example,
Noun - word, girl
Adverb - Now, There

Closed Class

① It has relatively fixed membership.
② It is also called as Function words.
③ It includes pronoun, article, preposition, auxilary, conjuction.
④ Here, words are finite in number.
⑤ They are mutually exclusive.
⑥ Example,
Conjuction - And, or, but
Preposition - at, in, on

Parsing

- Parsing in NLP is the process of determining the syntactic structure of a text by analyzing its constituent words based on an underlying grammar
- Parsing is the process of taking a string and a grammar and returning a parse tree.
- Parsing can be viewed as a search problem.
- Types of Parsing
① Top Down Parsing
② Bottom Up Parsing

Top-Down Parsing

- Top down parsing is goal oriented
- It is a parsing strategy that first looks at the higher level of the parse tree and works down the parse tree by using the rules of grammar.
- Top down parsing attempts to find the left most derivations for an input string.
- In this parsing technique, we start parsing from the top (start symbol) to down (leaf node) in a top-down manner.
- This parsing technique uses left most derivation.
- The main leftmost decision is to select what production rule to use in order to construct the string.
- Example: Recursive Descent Parser.

Bottom-Up Parsing

- Bottom-up parsing is data directed.
- It is a parsing strategy that first looks at the lowest level of the parse tree and works up the parse tree by using the rules of grammar.
- Bottom-up parsing can be defined as an attempt to reduce the input string to the start symbol of a grammar.
- In this parsing technique, we start parsing from the bottom (leaf node) to up (start symbol) in a bottom-up manner.
- This parsing technique uses right most derivation.
- The main decision is to select when to use a production rule to reduce the string to get the starting symbol.
- Examples: Shift Reduce Parser.

Module 4
Semantic Analysis

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 is used to train machine learning models.
   - Only problem is that such corpora are difficult to create.
3. Semi-supervised methods
   - Due to lack of such corpora, most WSD algorithms use semi-supervised methods.
   - The process starts with small amount of data which is also called as seed data.
4. Unsupervised method
   - This is the greater challenge to researchers and NLP professionals.
   - A key assumption is that similar meaning and senses occur in a similar context. They are not dependent on manual efforts.

WSD Evaluation

- Evaluation of WSD requires 2 inputs.

1. A Dictionary
   - First input of WSD evaluation.
   - It is used to specify the senses to be disambiguated.
2. Test Corpus
   - Another input of 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

Dictionary (Knowledge) Based Approach

- It is a simple approach to segment text is to scan each character one at a time from left to right and look up those characters in a dictionary.
- If the series of characters found in the dictionary then we have a matched word and segment that sequence as a word.
- But this will match a shorter length word.
- There are several ways to better implement this approach.

① Maximal matching
② Bi-directional maximal matching
③ Maximum matching

---

① Maximal matching
- One way to avoid matching the shortest word is to seek out longest sequence of characters within the dictionary.
- This approach is named as the longest matching algorithm or maximal matching. This is often a greedy algorithm that matches the longest word.

② Bi-directional maximal matching
- One way to solve this issue is to match backward.
- This is named as bi-directional maximal matching.
- It goes from left to right i.e. forward matching first and then from the end of the sentence. It goes from right to left. i.e. Backward matching
- Then it chooses the best result.

③ Maximum Matching
- One approach to solve the greedy nature of longest matching is an algorithm called 'maximum matching'.
- This approach segments multiple possible combinations and chooses the one with fewer words in the sentence.
- It also prioritizes fewer unknown words.

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
- 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 synsets.
- For example, the synsets "stones" and "cement" belong to synset "Building materials" or the synset "stones" also belong to another synset called "stonework."
- In above example, stones and cement are called hyponyms of synset building materials and also synsets building materials and stonework are called synonyms.
- Every member of a synset denotes same concept but not all synset members are interchangeable in context.
- There are three principles the synset construction process must adhere to
  ① minimality
  ② Coverage
  ③ Replaceability
- WordNet is similar to a dictionary, in that it groups words based on their meanings.
- However, there are some imp differences
  ① WordNet interlinks not only word forms or strings of letter but particularly senses of words. As a result, words that are found near one another are semantically clarified.
  ② WordNet marks semantic relations among words, whereas the grouping of words in a dictionary does not follow any particular pattern other than meaning similarity.

Senses

- One word can have multiple meanings
- The unit of meaning is a sense.
- A word sense is a discrete representation of one aspect of the meaning of a word

Relations between words/senses.

① Hyponymy and Hypernymy

- It refers to a relationship between a general term and the more specific terms that fall under the category of the general term.
- For example, the colors red, green, blue, yellow are hyponyms. They fall under general term of colour, which is the hypernym.

② Synonymy

- It refers to words that are pronounced and spelled differently but contain the same meaning.
- Example: Happy, joyful, glad.

③ Antonymy

- It refers to words that are related by having opposite meanings to each other.
- There are 3 types of antonyms
  - (a) Graded antonyms
  - (b) Complementary antonyms
  - (c) Relational antonyms.
- Example:
  - dead x alive
  - long x short

④ Homonymy

- It refers to the relationship between words that are spelled or pronounced the same way but hold different meanings
- Example,
  - bank (of river)
  - bank (financial institution)

⑤ Polysemy

- It refers to a word having two or more related meanings.
- Example,
  - bright (shining)
  - bright (intelligent)

⑥ Meronymy

- It is a logical arrangement of text and words that represent a part of or member of something.
- Example: A segment of an apple.

Module 5
Pragmatics

Reference Phenomenon

- The set of referential phenomenon that natural language provide is quite rich indeed.
- There are five types of referring expressions.

① Indefinite Noun Phrase
② Definite Noun Phrase
③ Pronouns
④ Demonstratives
⑤ Names

① Indefinite Noun Phrase
- Such kind of reference represents the entities that are new to the hearer into the discourse context.
- For example, In the sentence -
"Ram had gone around one day to bring him some food."
Here, some is an indefinite reference.

② Definite Noun Phrase
- It is opposite to Indefinite Noun phrase.
- Such kind of reference represents the entities that are not new or identifiable to the hearer into the discourse context.
- For example, in the sentence,
"I used to read The Times of India".
Here, The Times of India is definite reference.

③ Pronouns
- It is a form of definite reference
- For example, "Ram laughed as loud as he could."
Here, he represents pronoun referring expression

④ Demonstratives
- These demonstrate and behave different than simple definite pronoun.
- For example, this and that are demonstrative pronoun

⑤ Names
- It is simplest type of referring expression
- It can be the name of a person, organization or location. Ex. Ram

Module 6
Applications (Preferably for Indian Regional Languages)

Machine Translation
- It is also known as automated translation.
- Machine Translation (MT) is simply a procedure when a computer software translates text from one language to another without human assistance.
- At its fundamental level, machine translation performs a straightforward replacement of atomic words in a single characteristic language for words in another.
- In simple terms, we can say that machine translation works by using computer software to translate the text from one source language to another target language.
- Thus, Machine Translation is the task of automatically converting one natural language into another, preserving the meaning of the input text and producing fluent text in the output language.

Challenges of Machine Translation
1. The large variety of languages, alphabets and grammars.
2. The task to translate a sequence to sequence is harder for a computer than working with number only
3. There is no one correct answer.

Types of Machine Translations.
1. Statistical machine Translation (SMT)
- SMT works by referring to statistical models that are based on the analysis of large volumes of bilingual text.
- It aims to determine the correspondence between a word from the source language and a word from the target language.
- Example, Google Translate
- SMT is great for basic translation but it's biggest drawback is that it does not factor in context, i.e. Translation is of poor quality

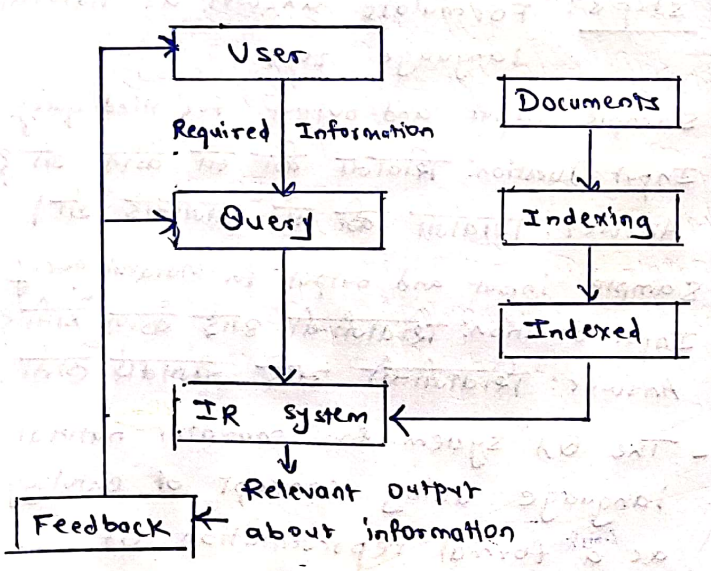
2. Rule-Based Machine Translation (RBMT)
- RBMT translates on the basis of grammatical rules.
- It conducts a grammatical analysis of source language and target language to generate the translated sentence
- But, RBMT requires extensive proofreading and its heavy dependence on lexicons means that efficiency is achieved after a long period of time.

3. Hybrid Machine Translation (HMT)
- It is a blend of SMT and RBMT.
- It uses a translation memory, making it far more effective in terms of quality
- Drawback of HMT is the need for extensive editing and human translator will be required.

4. Neural machine Translation (NMT)
- It is a type of machine translation that depends on neural network models to develop statistical models for the purpose of translation
- Benefit of NMT is that it provides a single system that can be trained to decipher source and target text.

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 information.
- 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 example, 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. Documents
  2. Queries
  3. Matching Function that compares queries to documents.
- Types of IR model
  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 operation.
     - 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)

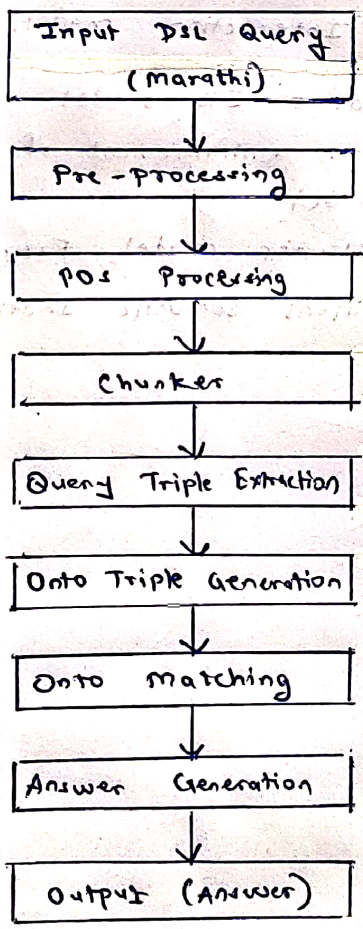
Information Retrieval	Information Extraction
① Document Retrieval	① Feature Retrieval
② Return set of relevant documents	② Return facts out of documents
③ The goal is to find documents that are relevant to the user's information need.	③ The goal is to extract pre-specified features from documents
④ Real information is buried inside documents.	④ Extract information from within the documents
⑤ The long listing of documents	⑤ Aggregate over the entire set.
⑥ Used in many search engines - Google is the best IR system for the web	⑥ Used in database systems to enter extracted features automatically
⑦ Typically uses a bag of words model of the source text.	⑦ Typically based on some form of semantic analysis of the source text.
⑧ Mostly use the theory of information, probability and statistics.	⑧ Emerged from research into rule-based system.

Application considering Indian regional language

Question Answering System

- It is a branch of learning of Information Retrieval and NLP.
- Question Answering focuses on building systems that automatically answer questions posed by humans in natural language.
- A computer understanding of natural language consists of the capability of a program system to translate sentences into an internal representation so that this system generates valid answers to questions asked by a user.
- Valid answer means answer relevant to questions posed by the user.
- Example.

Question Answering System for Hindi and Marathi language



Question Answering System

Algorithm

Input: Natural language question in Hindi / Marathi language.

Output: Answer in Hindi/Marathi language.

step 1: Tokenize the input question into word tokens.

step 2: Group the correlated words into one merged word.

step 3: Extract POS tag of each word in the tokenized list.

step 4: Chunk the POS tagged words into noun and verb groups

step 5: Extract query triple from chunked grouped list.

step 6: Generate onto triple

step 7: Traverse ontology to fetch answer.

step 8: Formulate answer as natural language text.

Sample input and output for Hindi query

Input Question: शिवाजी की माँ कौन थी?

Answer: शिवाजी की माँ जीजाबाई थी।

Sample input and output for Marathi query

Input Question: शिवाजीची आई कोण होती?

Answer: शिवाजीची आई जीजाबाई होती.

- The QA system for marathi natural language using concept of ontology as a formal representation of knowledge base for extracting answers
- Ontology is used to express domain specific knowledge about semantic relations and restricts in the given domains.
- The ontologies are developed with the help of domain experts and the query is analyzed both syntactically and semantically.
- The results obtained are accurate enough to satisfy the query raised by the user.
- The level of accuracy is enhanced since the query is semantically analyzed.
- QA system has become part of daily life of users, over a period of time many personal assistance like Google, Siri, Cortana are developed which provide precise and accurate answer to user question