Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

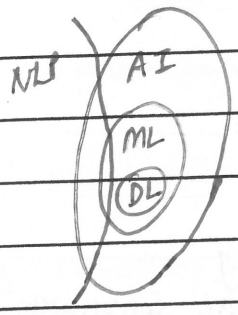
Subject: Natural Language Processing

Introduction to NLP : Origin & History of NLP, Knowledge of grammar, stages in NLP, Ambiguities and its types in English, Generic NLP system, Challenges of NLP, Applications of NLP.

Origin and History of NLP:

Natural language processing (NLP) is a part of everyday life. Language is the primary means of communication used by humans. It is a tool used to express ideas and emotions. Natural language processing is concerned with the development of computational models of aspects of human language processing.

Linguistic structure:



Definition:

It is a subfield of computer science in AI concerned with interactions between computers and human language enables the machine to interpret and understand the way humans communicate.

→ Why NLP? (Need of NLP).

→ • Unstructured data
• Multilingual websites.

History of NLP: Timeline.

- 1949 (Weaver's memorandum)
- 1957 (Chomsky Generative Grammar)
- 1970 (conceptual ontologies)
- 1980 (symbolic approach)
- 1990 (statistical approach).

CR.

ERA of NLP:

- Machine Translation (1940–1960)
- Flavoured with AI (1960–1980)
- 1980– current.

Phase I: (1940–1960)

- Research on NLP started with machine translation.
- Automatic machine translation from Russian to English.
- It was very basic and rudimentary

Phase2: second phase work was AI flavoured.

- Baseball Question Answering system (1961)
- Emphasis on world Knowledge.
- Manipulation of meaning representations.

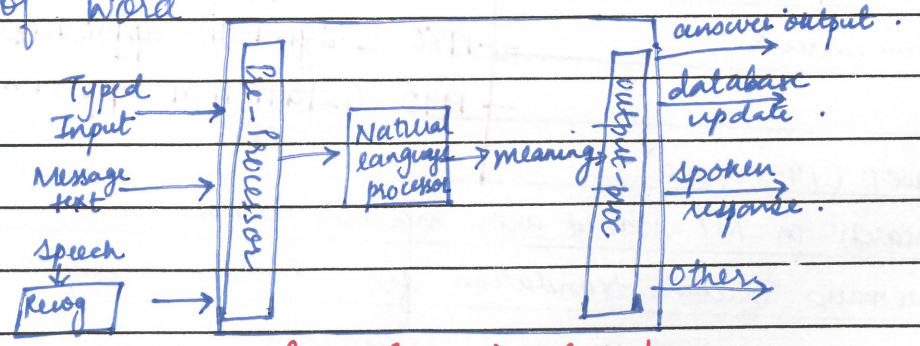
Phase3: SHRDLU (1973) an early natural language understanding computer program, developed by Terry winograd at MIT in 1968-1970.

- Woods (1978) answering questions about ATN grammars.
- Arpa speech understanding project (1980).
- Front end projects were used for interfacing with large automatic databases. Intellect was the first commercial front end.

Phase 4: 1980-2000

- Statistical machine translation
- Rule based machine translation
- Computational Grammar Theory.

Generic NLP System: Any natural language processing system should start with some input and end with effective and accurate output. The inputs for Natural language processor can be text or speech. The outputs that can be generated by the system could be form of answer when the input is question. outputs can be database update spoken response, semantics, part of speech or semantics of word



fig(1) Generic NLP system.

Levels of NLP : (Imp)

Natural language processing works on multiple levels and most often these different areas synergize well with each other. The NLP can be broadly divided into different levels

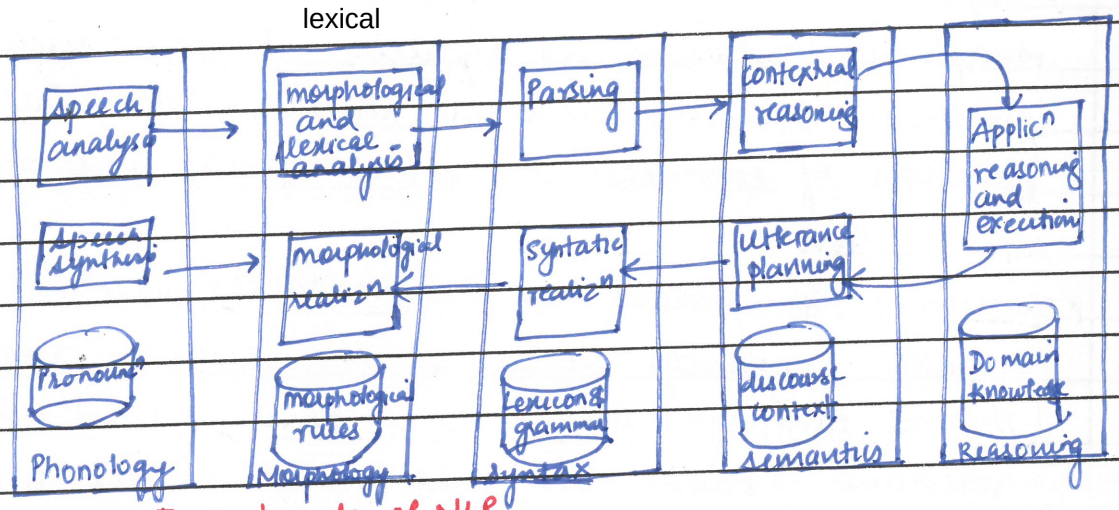


Fig 2: Levels of NLP.

Phonology: It deals with interpretation of speech sound within and across words.

Morphology: It is a study of the way words are built up from smaller meaning bearing units called morphemes. for example fox has a single morpheme while 'cats' has two morphemes 'cat' and morpheme '-s' representing singular and plural concepts. Morphological lexicon is the list of stem and affixes together with basic information.

Syntax: It is the study of formal relationships between the words. It is the study of how words are clustered in form of classes in POS (part of speech), how they are grouped with their neighbours into phrases and the way words depend on each other in a sentence.

Semantics: It is the study of meaning of words that are associated

(4)

with grammatical structure. It consist of two types of approaches syntax driven semantic analysis and semantic grammar.
In discourse context the level of NLP works with the text longer than a sentence. There are two types of discourse: anaphora resolution and discourse/text structure recognition. Anaphora resolution is replacing of words such as pronouns. Discourse structure recognition determines the function of sentences in a text which adds meaningful representation.

Reasoning: To produce an answer to the question which is not explicitly stored in the database, NLIDB (Natural language Interface to database) carries out the reasoning based on data stored in the database. For example, consider a database that holds academic information about the student and user query posted such as: 'which student is likely to fail in science subject'? To answer this NLIDB needs a domain expert to narrow down the reasoning process.

Stages in NLP (Imp) (MU 2023 (Dec), 2024 (May)) (10M).

1. Lexical Analysis
2. Syntatical Analysis
3. semantic analysis
4. discourse analysis
5. Pragmatic analysis

There are five stages of NLP namely lexical analysis, syntatical analysis, semantic analysis, discourse analysis and pragmatic analysis.

Lexical Analysis: 1] Lexical Analysis is the first stage of NLP, it is also known as morphological analysis.

2] At this stage the structure of words is identified and analyzed.

3] Lexicon of the ~~word~~ language means collection of words and phrases in a language.

4] Lexical Analysis is dividing the whole position of text into paragraphs, sentences and words.

Syntatical Analysis 1] This involves analysis of words in a sentence for grammar and ordering the words, in a way that it potrays the relationship among words.

eg> The sentence such as " The school goes to girl" is rejected by English syntatical analyser.

Semantic Analysis: 1] Semantic Analysis draws the exact meaning or dictionary meaning from the text.

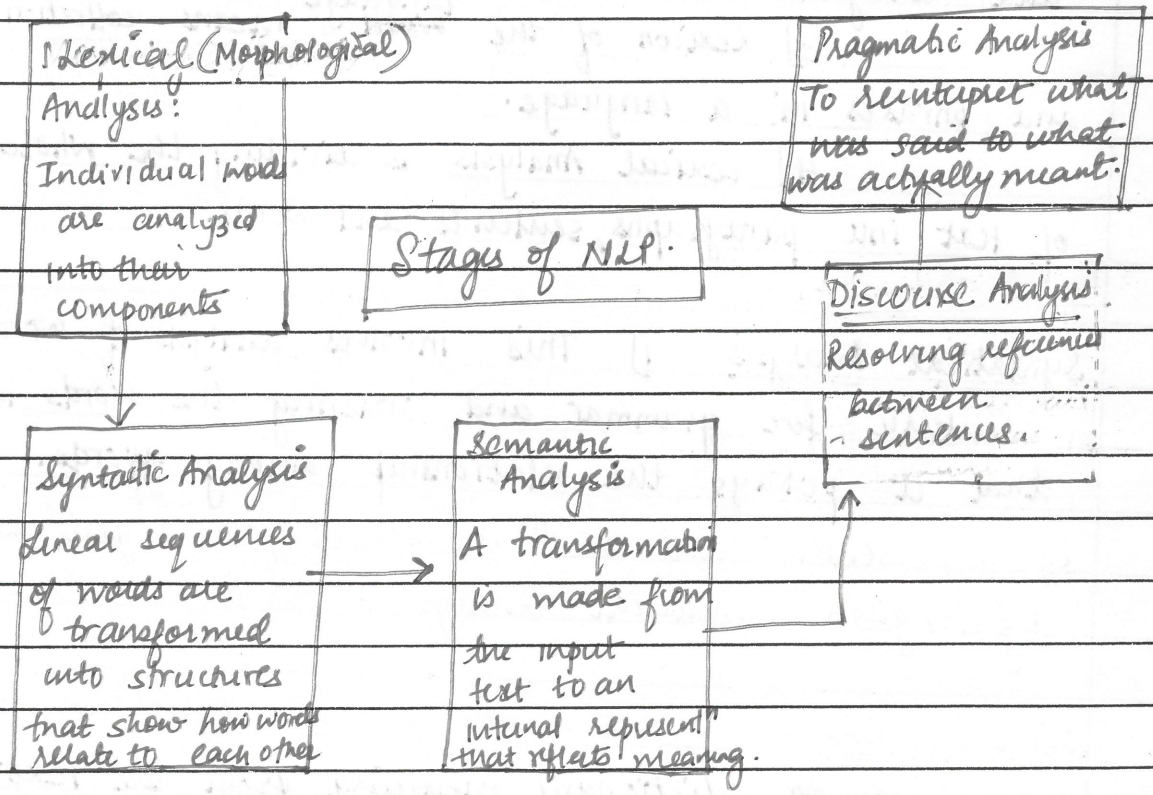
2] The text is checked for meaningfulness. It is done by mapping syntatical structures and objects in the task domain.

3] The semantic analyser neglects sentence such as "Hot ice-cream".

Discourse Integration : 1] The meaning of any sentence depends upon the meaning of sentence before it. Furthermore it brings about the meaning of immediately following sentence.

For example: Meena is a girl, she goes to school here "she" is a dependency pointing to Meena.

Pragmatic Analysis: (1) This concerns how sentences are used in different situations and how it affects the interpretation of the sentence. It involves deriving those aspects of language which require real world knowledge.



Morphological / Lexical Analysis: The morphological level of linguistic processing deals with study of word structures and word formation. The most important unit of morphology defined as having 'minimal unit of meaning' is referred to as morphemes. For example the word "unhappiness" it can be broken down into three morphemes (prefix, stem, affix) and each convey some meaning the prefix un- refers to not being" while suffix "ness" refers to the state of being, the stem "happy" is considered as a

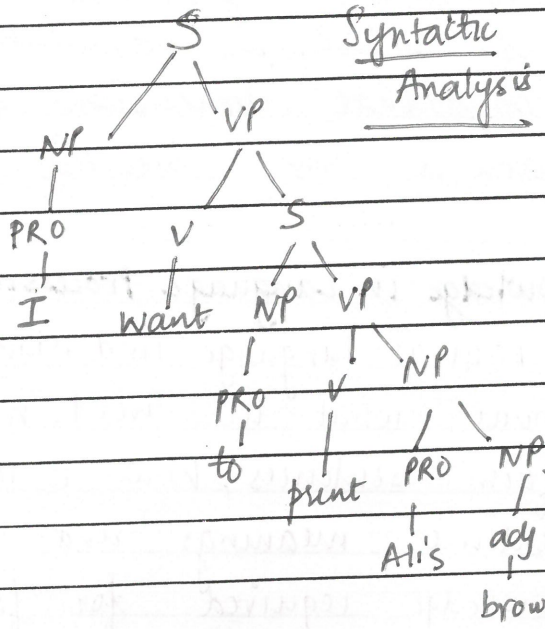
free morpheme. Bound morpheme require a free morphemes to which it can be attached to and therefore not appear as a word of their own.

In information retreival, the document and query terms can be stemmed to match the morphological variants of terms between documents and query.

Syntactic Analysis: The part of speech tagging output of lexical analysis can be used at the syntactic level of linguistic processing to group words into phrases. Syntactic analysis are also referred to as parsing. Parsing allows extraction of phrases which convey more meaning than just individual words by themselves, such as noun phrase.

In information retreival parsing can be leveraged to improve indexing

I (pronoun)
want (verb)
to (prep)
print (verb)
Nikhil (noun)
is (poss)
brown (adj)
file (noun)



Semantic Analysis:

The semantic level of linguistic processing deals with what a sentence really means and disambiguating words with multiple definitions to the given context. This level entails appropriate interpretation of the

meaning of sentences, rather than analysis at a level of individual words or phrases.

Discourse Analysis:
The discourse level of linguistic processing deals with analysis of structure and meaning of text beyond a single sentence, making connections between words and sentences. At this level, Anaphora resolution is also achieved by identifying entity referenced by an anaphor

"I voted for Obama because he was most aligned with my values," she said.

Pragmatic Analysis: The pragmatic level of linguistic processing deals with the use of real world knowledge and understanding how this impacts the meaning of what is communicated. This analyses the contextual dimension of the documents and query and a more detailed representation is derived.

→ **Knowledge in Language Processing.**
A natural language understanding system must have knowledge about what the words mean, how words combine to form sentences, how word meanings combine to form sentence meanings and so on. The different forms of knowledge required for ~~form~~ natural language understanding are given below:

Phonetic and Phonological knowledge: Phonetics is a study of language at the level of sounds while phonology is the study of the combination of sounds into organized units of speech the formation of syllables and larger units. Phonetic and phonological knowledge are essential for speech based systems as they deal with how words are related to the sounds that realize them.

Morphological knowledge: Morphology concerns word formation It is a study of patterns of formation of words by combination of sounds into minimal distinctive units of meaning called as morphemes. Morphological knowledge concerns how words are constructed from morphemes.

Syntactic knowledge: Syntax is the level at which we study how words combine to form phrases, phrases combine to form clauses and clauses combine to form sentences. It deals with how words can be put together to form correct sentences. It also determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases.

Semantic knowledge: It concerns the meanings of the words and sentences. This is study of context independent meaning, no matter in which context it is used. Defining the meaning of a sentence is very difficult due to ambiguities involved.

Pragmatic knowledge: Pragmatics is the extension of the meaning or semantics. Pragmatics deals with the contextual aspects of meanings in particular situations. It concerns how sentences are

used in different situations and how the use affects the interpretation of the sentence.

Discourse Knowledge: Discourse concerns the connected sentences. It is the study of chunks of language which are bigger than a single sentence. Discourse concerns inter sentential links that is how the immediately preceding sentences affect the interpretation of the next sentence. Discourse knowledge is important for interpreting pronouns and temporal aspects of the information conveyed.

World knowledge: World knowledge is nothing but everyday knowledge that the speakers share about the world. It includes the general knowledge about the structure of world and what each language user must know about other users belief and goals.

Knowledge representation and reasoning systems have incorporated natural language as interfaces to expert systems or knowledge bases that performed tasks separate from natural language processing.

Ambiguity in NLP (MU 2022(Dec), 2023(May), CMPN). (Imp).

Natural language has a very rich form and structure. It is very ambiguous. ~~Ambio~~ Ambiguity in a language is the ability of being understood in more than one way. Natural language is very ambiguous. NLP has following ambiguities:-

- Lexical Ambiguity
- Syntactic Ambiguity
- Semantic Ambiguity
- Anaphoric Ambiguity
- Pragmatic Ambiguity

Ambiguity can occur at all levels. It is a property of linguistic expressions. If an expression (word/phrase/sentence) has more than one interpretation we can refer it as ambiguous.

for example: Consider the ~~sitem~~ sentence, "The chicken is ready to eat"
Interpretations can be: · The chicken (bird) is ready to be fed.
or · The chicken (food) is ready to be eaten.

Consider another sentence " There was not a single man at the party" The interpretations can be
· Lack of Bachelors at a party or
· Lack of men altogether.

There are different types of ambiguities:—

1. Lexical Ambiguity: This is the ambiguity of a single word. A word can be ambiguous with respect to its syntactic class. Eg book, study.
for eg: The word "silver" can be used as a noun, adjective or a verb.
- She bagged two silver medals.
- She made a silver speech
- Her worries had silvered her hair.

Lexical ambiguity can be resolved by lexical category disambiguation ie part of speech tagging. As many words may belong to more than one lexical category part of speech tagging is a process of assigning part of speech or lexical category such as a noun, verb, pronoun to each word in a sentence.

Lexical semantic ambiguity: This type of ambiguity occurs when a single word is associated with multiple senses. Eg bank, pen, fast, bat, cricket etc. eg) The tank was full of water.
· I saw a bat.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

Possible meaning of words bat = flying mammal / wooden club
saw = past tense of see / present tense of saw.

The occurence of 'tank' corresponds to the syntactic category
Lexical semantic ambiguity can be resolved using word sense disambiguation.

2 **Syntactic Ambiguity:** The structural / syntactic ambiguities are classified into two types (1) Scope Ambiguity
(2) Attachment Ambiguity

i **Scope Ambiguity:** This involves operators and quantifiers.
Consider the example: Old men and women were taken to safe locations
The scope is highly ambiguous, (old men and women) or ((old men) and women), The scope is often not clear and this creates ambiguity.

ii **Attachment Ambiguity:** A sentence has attachment ambiguity if the constituent fits more than one position in a parse tree. Consider the example:
The man saw a girl with a telescope.
It is ambiguous as to whether a man saw a girl carrying a telescope or he saw her through the telescope. The meaning is dependent on whether the preposition 'with' is attached to the girl or man.

Eg 2) Buy books for children.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

3. **Semantic Ambiguity:** This occurs when the meanings of the words themselves can be misinterpreted. Even after the syntax and the meanings of the individual words have been resolved consider the example: "Seema loves her mother and Sriya does too"
Interpretations: Sriya loves seema's mother or Sriya likes her own mother
Consider another example: "The car hit the pole while it was moving" the interpretations can be:
- The car while moving hit the pole.
- The car hit the pole while the pole was moving"

4 **Discourse Ambiguity:** Discourse level processing needs a shared world or shared knowledge and the interpretation is carried out using this context. Anaphoric Ambiguity comes under discourse level.

Anaphoric Ambiguity: Anaphora are entities that have been previously introduced into the discourse.
Consider the example: The horse ran up the hill. It was very steep. It soon got tired. (The anaphoric reference of it in two situations cause the ambiguity. steep applies to the surface hence 'it' can be a hill. Tired applies to animate object hence it can be a horse.)

5 **Pragmatic Ambiguity:** Pragmatic ambiguity refers to the situation where a context of a phrase gives it multiple interpretation. This is one of hardest task in NLP. This problem involves processing user intention, sentiment, belief world, modals etc. all of which are complex tasks. Consider the example:

66

Tourist (checking out of the hotel): waiter, go upstairs to my room and see if my sandals are there; do not be late; I have to catch the train in 15 minutes.

Waiter (running upstairs and coming back panting): Yes sir, they are there. (Clearly the waiter is falling in short of tourist expat's ~~[illegible]~~ He did not get the pragmatics of the situation.

Pragmatic ambiguity arises when the statement is not specific and the context does not clearly provide the information to clarify the statement. Informⁿ is missing and must be inferred. Consider the example:

"I love you too" this can be interpreted as

- I love you (just like you love me)
- I love you (just like someone else does)
- I love you (and I love someone else)
- I love you (as well as liking you).

** It is a highly complex task to resolve all these kinds of ambiguities, especially in the upper levels of NLP.

Statistical Approaches of Ambiguity resolution in Natural language Processing are 1] Probabilistic model

2] Part of speech tagging (Rule based Approaches)

3] Markov model approaches

4] Maximum Entropy approaches

5] HMM Based Taggers

Machine learning approaches.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

Challenges of NLP.

NLP is a powerful tool with enormous benefits still there are numerous natural language processing challenges:-

1. Contextual words, phrases and Homonyms: The same words, phrases can have diverse meanings according to the context of a sentence and many words have the exact same pronunciation but completely different meanings.
2. Synonyms: Synonyms can cause issues like contextual understanding since we use different words to express identical idea. Additionally some of these words may convey exactly the same meaning while some may be levels of complexity for example some, little, tiny, minute may have the same meaning
3. Irony and Sarcasm: Irony and sarcasm may present problems for machine learning models since they use words and phrases that strictly be positive or negative but truly mean opposite.
4. Ambiguity: Ambiguity in NLP refers to sentences and phrases with multiple interpretations, it is a highly challenging task to resolve all ambiguities. An NLP system must be able to seek the context and help to understand the phrasing. It may also need to ask the user for clarity.
5. Errors in Text, speech: Misspelled words or misused words can generate problems for text analysis. There is a need to use a NLP tool with capabilities to recognise common misspellings of words and move beyond them.
6. Language differences: Different languages can have not only different

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

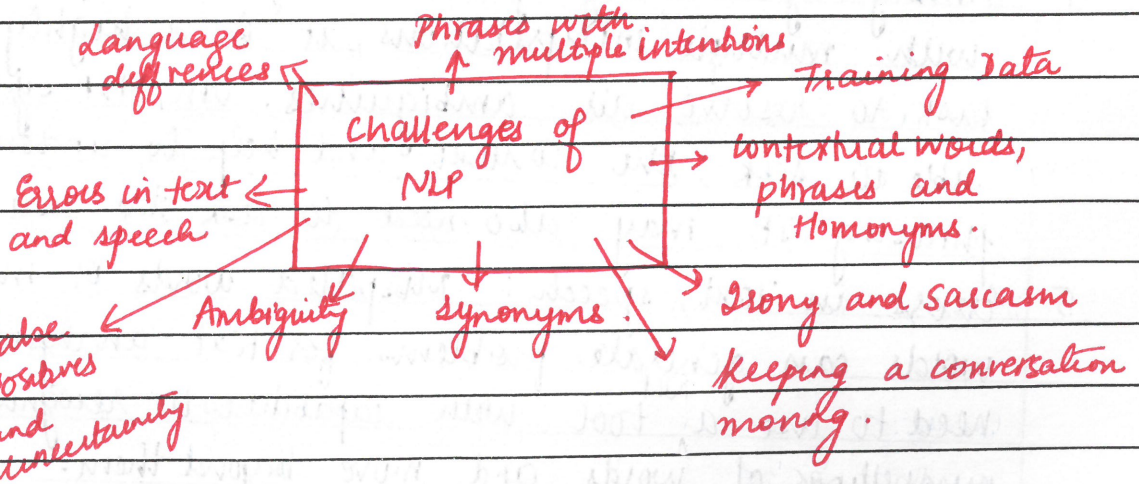
sets of vocabulary but also different types of phrasing, different modes of inflection and different cultural expectations.

7. Phrases with multiple intentions: Some phrases and questions actually have multiple intentions so NLP system cannot simplify the situation by only interpreting one of the intentions.

8. Training Data: NLP is all about analysing the language. The abilities of a NLP system depends on the training data provided to it.

9. False positives and uncertainty: A false positive occurs when a NLP notices a phrase that should be understandable but cannot be sufficiently answered. The solution here is to develop a NLP system that can recognise its own limitations.

10. Keeping a conversation moving: Many modern NLP applications are built on the dialogue between human and machine, hence the challenge is develop an interactive NLP system.



Applications of NLP (Nov 22, Dec 23, May 22) (CMPN) (V.V. Imp).

There are humongous applications of NLP. Natural language processing, machine learning and artificial Intelligence are used Interchangeably. AI is regarded as an umbrella term for machines that can simulate human intelligenic NLP and ML are regarded as subsets of AI. Natural language processing is a form of AI that gives machine the ability to not just read but to understand and interpret the human language. Some of the applications of NLP are as follows:—

1. Machine Translation 2. Speech Recognition 3. Speech Synthesis 4. Information Retreival 5. Information Extraction 6. Question Answering 7. Text summarisn 8. Sentiment Analysis.

1] Machine Translation: Machine Translation is used to translate text or speech from one natural language to another natural language. Translating languages is more complex task than a simple word to word replacement. For puforming translation, it is essential to have knowledge of words, phrases and grammar. (eg) Google translate.

2] Speech Recognition: (1) Speech recognition is the process where acoustic speech signals are mapped to the set of words. As there is a wide variation in the pronounciation of word, homonym. It uses NLP to allow computers to collect human interaction and ML to respond in a way that copies human responses.

3] Sentiment Analysis: (1) Sentiment Analysis uses NLP to interpret and analyze emotions in a subjective data like news articles and tweets. Positive, negative and neutral opinions can be identified to determine a customer's sentiment towards a brand, product and service. Sentiment Analysis is used to measure public opinion, monitor brand reputation and understand customer experiences.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

4] Information Retrieval: (IR). It is a scientific discipline that deals with analysis, design and implementation of a computerised system that addresses representation, organization and access to large amount of heterogeneous information encoded in a digital format. The search engine is a well known application of IR. The research area of information retrieval includes: information searching, information extraction, information categoriz^n, information summarisation.

5] Question Answering system: ① Question Answering system are intelligent systems that can provide answers to customer queries. Question answering systems have a huge array of knowledge and good array of language understanding. Question answering systems can be found in social media chats and tools such as Siri and IBM's Watson.

6] Automated Text Summarisation: ① Automatic text summaris^n is a task of condensing a piece of text to a shorter version. It extracts the main idea and preserves the meaning of content. This application of NLP is used in web search and bulletins of market reports.

7] Speech Synthesis: Speech understanding system attempts to perform a semantic and pragmatic processing of spoken utterance to understand what the user is saying and act on what is being said.

8] Information Extraction: It includes extraction of structured information from unstructured text. It is an activity of filling predefined template from natural language text. The research area includes named entity, resolving anaphora and identifying relationships between entities.