Artificial / Intelligence :-

↓

"Man" made + "Thinking Power".

It is branch of Computer Science by which we can create intelligent machine, which can behave like a human, think like humans & able to make decision.

→ AI do not need to Pre Program machine to do some work.

→ You have to create a machine with Programmed algorithm. which can work with own intelligence.

NLP :-

machine / Computer (AI) know only binary code (010:...), they don't know the human language that achieved by using NLP

code
↓
binary code
0 - low voltage
1 - high voltage
logic gate.

→ Why we study natural language Processing?

→ NLP stands for natural language Processing.

→ It is a subfield of AI.

→ It deals with the interaction b/w humans & computers in natural language.

→ It uses computational techniques to Process & analyse the data.

Applications:-

→ Question answering.

ex:- Alexa, Siri

→ Spam detection.

ex:- If you get a mail, that mail is danger. Spam.

→ Machine translation.

ex:- convert to one language to another.

→ Speech Recognition.

ex:- google assist, If you have youtube, that video has no subtitles it will click [CC] it shows text.

→ chatbot 🗨 chat box in website.

→ Sentiment analysis.

ex:- sad, happy, crying.

ex:- I am enjoying right now,

Phases of NLP :- 5 phases in NLP

1. Lexical analysis (or) Morphological analysis → converting the given sentence into stream of lexemes / morphemes

[Lexeme :- You giving sentence may be sub word / words]

[morpheme :- The smallest unit of words it has meaning]

2. Syntatic analysis → It is used to check grammar, word arrangements and relationships b/w words.

3. Sematic analysis (meaning of words / sentence / phases). → It is used to find the meaning of words, phases and sentences.

4. Discourse integration (group of sentences / paragraphs which are integrated together so that each sentences depends on the previous sentences)

ex:- (Ramu) is a standard, (1st)
(He) (Ramu) is a good boy (2nd)

↓

→ Integrating the sentences in such a way that the meaning of the sentence is effected by the preceeding sentence.

↓

5. Pragmatic Analysis → It first understands what is stated & then understands the intended meaning of the sentence.

(1)

words and their components:-

words:- words are nothing but a meaningful unit of a sentence, by using the meaningful words only we construct the meaningful sentence.

→ we have 4 components.

1. tokens
2. Lexemes
3. morphemes
4. typology.

tokens:-

It refers to a sequence of characters or units that are treated as a single entity.

→ Tokens are words that are created by dividing the text into smaller units.

→ **Process:-** Tokenization. (By using tokenization we obtained the tokens)

(1) word tokens

(2) Characters tokens.

(3) Subword tokens.

Ex:- I | love | reading | news papers.

↳ words tokens

r|e|a|d|i|n|g

↳ Character tokens.

news | Paper |

↳

Subword tokens

2. Lexemes:-

→ They are the base (or) canonical form of words

Ex:- The base word of running, ran is 'run' (base word)
inflection/variation

Ex2:- largest, larger is large (base word).

→ fundamental unit of meaning associated with a word, regardless of the inflection (or) variations.

Process:- Lemmatization.

3. morphemes:-

→ They are the smallest meaniful unit.

→ Many words are formed by combining more than one morpheme.

Types:-

1. Free morpheme :- Ex:- agree
2. Bound morpheme :- Ex:- disagreement

Ex:- disagreement.

Process:- morphological process

4. Typology:-

→ Typology refers to the classification (or) categorization of language based on their structural (or) grammatical features.

→ Aim:- To identify common Pattern & variation across language to understand their Properties (or) relationships.

1. Isolating (or) Analytical language:-

→ have no (or) relatively few words that are comprised of more than one morpheme.

Ex:- Chinese, Thai

(In this kind of languages ~~these~~ these are very less words which are formed by one (or) more morphemes).

2. Synthetic languages:-

It is opposite of 1st language.
Combines more morphemes in one word.

3. Agglutinative language:-

morphemes associated with only a single function at a time. (Complex words to explain a simple topic).

Ex:- Korean, Japanese, Tamil, finnish.

4. Fusional languages:-

Ex:- bat < animal / cricket bat.

-feature (meaning) — Pre-morpheme ration highes than one.

Ex:- Arabic, Sanskrit, German etc.

(1)

Issues and challenges:-

There are 3 issues and challenges.

1. Irregularity
2. Ambiguity
3. Productivity

Irregularity:-

The phenomenon where certain words (or) words forms doesnot follow regular Patterns (or) rules interms of their morphology (or) syntax.

→ It is a challege for algorithms which follow Particular Patterns

(i) Irregular verbs & nouns:-

which doesnot follow standard Pattern (or) inflection.

ex:-

Past	Present	future	
choose	chose	chosen (choose)	} regular words.
bite	bit	bitten (bite)	
went		(go) Irregular words	

(ii) Exceptional Inflection:-

Comparative and superlative adjectives.

ex:- big, bigger, biggest
dark, darker, darkest
good, better, best & Irregular words.

(2) Ambiguity :- (The word which is having multiple meaning's, same spelling).

word forms that can be understood in multiple ways out of context.

→ word forms that look same but have diffinct function (or) meaning (homonyms).

→ Ambuguity arises in morphological Processing and language Processing.

(a) word sense ambiguity :- Meaning depending on the context in which they are used.

Ex :- bat < animal / cricket bat. } bank < money / river banks

(b) Parts of Speech ambiguity :- Different Parts of speech on their usage.

Ex :- I run (Verb)

He went for a run (noun).

(c) Structural ambiguity :- multiple valid syntactic structure.

(d) Referential Ambiguity :- (→ Pronoun)

Refering a Person by he/she

girl :- she / her, boy :- He / His.

(3) Productivity :-

Ability to generate new words (or) word forms using Productive rules. (or) generate some new words by taking some unknown words).

Ex:- According to wikipedia,

- google means 1 followed by 100 zero's

↓

Term this word, an unknown word [google] is generated.

↓ By Productivity rules.

new words are generated based on that word

i.e, googling, googlish, googleology etc.

→ Names of People, organisation, location have unique structure where the Productivity issues arises.

1)

Morphological Models:-

- This morphological models used to analyze the structure and also formation of the words.

→ It also helps us to identify the morphemes in a particular sentences.

5 morphological models.

1. Dictionary lookup.
2. finite-state morphology.
3. Unification-based morphology.
4. Functional morphology.
5. Morphology Induction

(a) Dictionary lookup:- This used to analyze the structure of a particular word.

word → base form (or) canonical form
Eg:- beautiful → beauty

Search in dictionary

retaining these information.

→ If you don't found the word in dictionary (or) you see more topics in that particular word. It will combine the another algorithm and shown the result word.

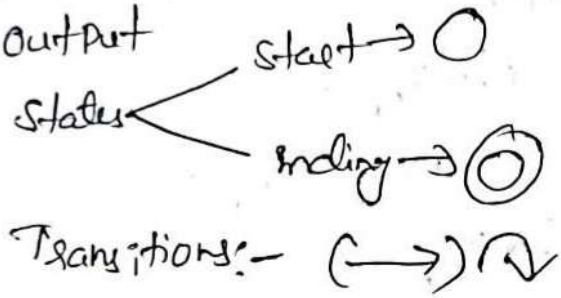
(b) Finite-state morphology:-

→ Based on finite-Automation and formal language theory.

→ used for generation & recognition tasks.

formal language theory:-

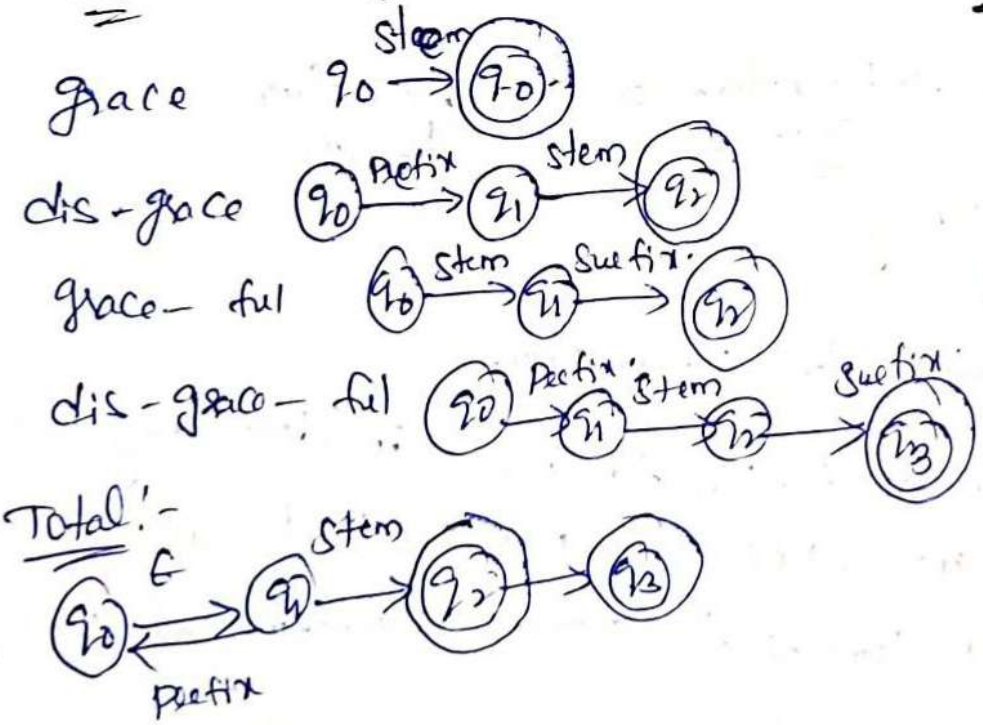
Input
output
States
Transitions:-



Finite-state transitions:-

This Process is represented by FST's

Ex:- word-grace



(c) Unification-based morphology:-

One word which is formed by combining three difference words. that formed word should be meaning ful.

word = word1 + word2 + word3.

↳ meaningful.

eg:- WWW = world + wide + web.

DIY = Do + it + yourself

HDTV = High + definition + Television.

(d) Functional-based morphology:-

→ used to analyze the role of a morpheme in a word. (which is helpful in forming another word).

→ and also analyze their contribution in overall grammatical structure of a sentance.

eg:- (walk), walking.

↓ word.

(e) morphological Induction:-

→ It is used to discover the Pattern of a particular word (or) structures without helping in any human being. The system only automatically detects.

→ By Processing classical morphology

10:55 - 11:45

Finding the structure of documents:-

→ Extracting the structure of documents helps Natural language Processing tasks like Parsing, machine translation etc.

How?

→ By chunking (Parsing) the input text (or) speech to blocks → segmentation.

→ Difference approaches for different languages.

For example:-

Chinese documents → (No white spaces).
→ 1. Segment Characters sequence into words.

Morphologically rich language. → 1. Preprocessing step (determines the tokens).
2. Apply algorithms.

→ In order to chunk the text, Segmentation is used.

Segmentation:- (Given text into some sentences).

→ Decides whether to mark a boundary b/w 2 tokens as sentence boundary (or) topic

2 types.

(i) Sentance- boundary detection.

(ii) Topic- boundary detection.

→ (i) Sentance - boundary detection :-

→ Also called Sentance Segmentation.

→ The Process of Segmenting sequence of words into units.

Sentance can be identified by :-

In written english :-

→ Beginning of the sentence (UPPER case).

→ End of sentence (?, ., !)

But Sometimes.

→ Capitalizations are ~~called~~ used for nouns and abbreviations.

→ Punctuation marks are used inside the sentences.

Example :-

I talked to Dr. Smith and my house is on mountain Dr.

(ii) Optical characters Recognition :- Segments the I/p text into sentence. These systems confuses with Periods and Commas.

→ It results in meaningful sentence.

Spoken character segmentation:-

(i) Automatic Speech Recognitions:- Segments the speech into sentence. These systems confuses with Punctuation marks (?, !, .)

→ code switching Problem arises when segmentation done for othe languages.

Example:- Spanish uses inverted quation marks & exclamating marks (¿ and ¡)

→ conventional rules-based Patterns are used to identify Potential ends of sentence.

(ii) TOPic boundary detection:-

→ It is also called as discourse segmentation/text segmentation

→ The Process of dividing the speech into homogenous blocks is called topic segmentation.

Applications:-

1) Text extraction and Retrieval.

2) Text summarigation.

Identification:-

(i) Text segmentation clues:-

→ by following the headlines.

→ By Paragraph breaks.

(ii) Speech segmentation clues:-

→ Pause duration

→ Speaker changes.

Methods:-

1) Generative Sequence classification Method:-

observation :- words, Punctuation. (?, !)

labels :- Sentence boundary, topic boundary.

→ These methods have models which are not only capable of Predicting the most likely label but also generating the sequence itself based on learned Probabilities.

One Example of this method is Hidden markov:-

How HMM ~~does~~ works?

1) Learns from data about the observations and its corresponding hidden state (POS).

2) Predict the labels for sequence generation is done.

3) classifies the new sequence.

Example:-

Sentence 1:- 'I/PRON Love/VERB coding/NOUN'

Sentence 2:- "The/DET quick/ADJ brown/ADJ fox/NOUN JUMPS/VERB".

Sentence 3:- 'She/PRON sells/VERB seashells/NOUN'.

→ Based on these 3 sentences sequence generation and classification is done.

2) Discriminative Local Classification method:-

Local features:- word identifies, Prefix, Suffixes & nearby POS (Part of Speeches).

→ These approaches focuses on making decision based on local information, typically without considering the entire sequence of data.

→ These methods directly model the relationship b/w input features and output labels.

Examples of these method be maximum Entropy markov model (or) support vector machine.

Applications:-

1) Parts-of-Speech Tagging. (To Predict the POS of the particular sequence)
2) Speech Recognition. (Ex:- Youtube)
3) Named Entity Recognition. (Person name, Location name etc).

Complexity of the approaches:-

* Whether comes to Complexity / one approach may be better than other.
* The complexity can be measured by rating their training and Prediction.
* Also measure their Performance.

Performances of the APPROACHES:-

$$\text{Error \%} = \frac{\text{no. of Errors}}{\text{no. of examples}}.$$

$$\text{Precision} = \frac{\text{total no of correct positive Predictions}}{\text{total no. of Positive Predictions.}}$$

$$\text{Recall} = \frac{\text{total no. of all correct Positive Prediction}}{\text{total no. of Positive instances.}}$$

$$\text{f1score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$