COMPUTER ENGINEERING DEPARTMENT

ASSIGNMENT NO. 1

Subject: Natural Language Processing

COURSE: B.E Year: 2021-2022 Semester: VIII
DEPT: Computer Engineering
SUBJECT CODE: DLO8012 SUBMISSION DATE: 06/03/2022

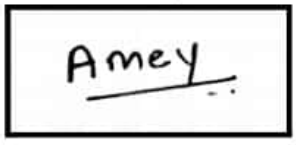
Roll No.: 50 Name: Amey Thakur

Class: BE-Comps B Date of Submission: 06/03/2022

NLP Assignment - 1

Sr. No.	Questions
1	What constitutes a generic NLP system?
2	Explain the various NLP processing steps.
3	Describe in detail the various ambiguities in NLP?
4	What is morphology? Explain with an example inflectional and derivational morphology.
5	What is FST (Finite state transducer)? What is their use in morphological analysis?
6	What is stemming and lemmatization and explain their uses in NLP.
7	What is RE (Regular expression)? Explain the use of RE in morphological analysis.
8	What is POS tagging? Explain the different types of POS tagging.
9	What is N-Gram? How is it used for Word sense disambiguation in NLP?

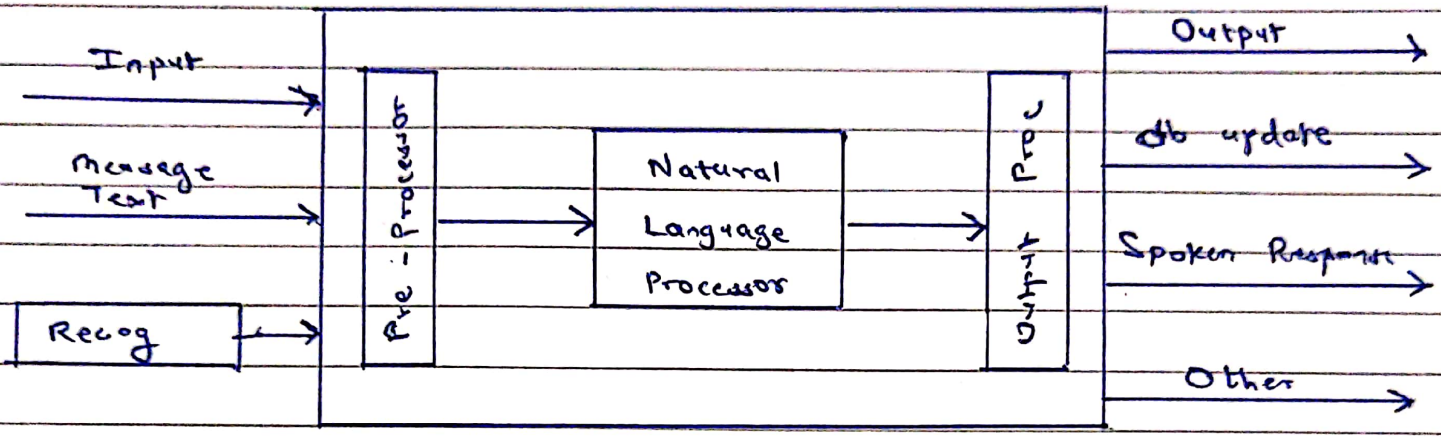
Student Signature:



AMEY THAKUR B-50 Amey.

Q1. What constitutes a generic NLP system?

Ans:



Generic NLP system

- Any natural language Processing should start with some input and end with an effective and accurate output.
- The input of NLP processor can be a text or speech
- There are variety of output that are generated by the system.
- Outputs can be a form of answer when input is a question
- Similarly, output can be a database update, spoken responses, semantics, Part of Speech, etc.
- A generic NLP system consists of:
  1. Recog
  2. Pre processor
  3. Natural Language Processor
  4. Output Proc.

AMEY THAKUR B-50 Amey.

Q2. Explain the various NLP processing steps.

Ans:

NLP Processing steps

① Lexical Analysis
- It is the first step in NLP.
- It is also known as morphological analysis.
- Here the structure of words is identified and analyzed.
- It divides the whole input text into sentences, words, Paragraphs.

② Syntactic Analysis
- It is also known as parsing.
- It involves word analysis to depict grammatical structure.
- The relation among these words are formed
- Eg. "The school goes to the girl" is rejected.

③ Semantic Analysis
- It draws exact meaning from text
- In this, text is checked for its meaningfullness.
- It is done by mapping syntactic analysis.
- Eg. "Hot icecream" is rejected.

④ Discourse Integration
- Sense of the context.
- Meaning of any single sentence depends on the sentence that precedes it and also invokes the meaning of the sentences that follows it.
- Eg. The word 'it' in the sentence "I knew it" depends upon the prior discourse text.

AMEY THAKUR B-50 Amey.

(5) Pragmatic Analysis

- In this, what is said earlier is re-interpreted on what is truly meant.
- Eg "John saw Mary in garden with the cat."
- Here we cannot predict whether John or Mary is with the cat.

AMEY THAKUR B-50 Amey

Q3. Describe in detail various ambiguities in NLP.

Ans:

Ambiguities in NLP.

① Lexical Ambiguity
- Where words have multiple assertions. This ambiguity takes place.
- Eg back stage (noun)
  back door (adjective)

② Syntactic Ambiguity
- It means sentences are parsed in multiple syntactical forms.
- Eg. "I saw a girl on the beach with binoculars."
  The confusion in the sequence leads to syntactic ambiguity

③ Semantic Ambiguity
- It is related to sentence interpretation
- Eg "I saw a girl on the beach with binoculars."
  Here meaning can be "I saw a girl through binoculars" or "The girl had binoculars"

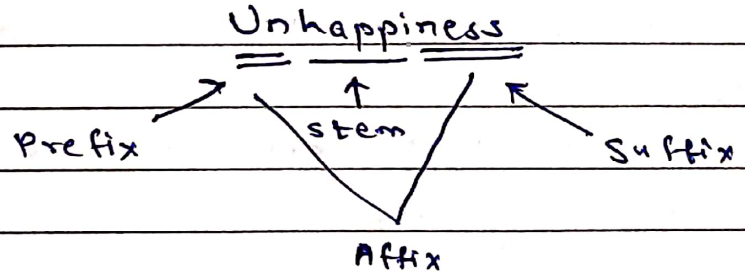
④ Metonymy Ambiguity
- It is the most difficult ambiguity. It deals with phrases in which the literal meaning is different from figurative assertion.
- Eg. "Nokia is screaming for new management".
  Here it really does not mean that the company is literally screaming

AMEY THAKUR B-50 Amey.

Q4. What is Morphology ?
Explain with an example inflectional and derivational.

Ans:

- NLP has various level and complexity of processing the data.
- One of the techniques of analyzing natural language text is to break it down in small constituents like words, sentences and paragraphs.
- It is important to understand the meaning of words before knowing the syntax because words are the fundamental units for processing.
- Morphology is nothing but the study of word formation from its minor meaning bearing unit morpheme.
- There are two types of morphemes.
  1. Stem
  2. Affix
- Example:



- The main morpheme is called as stem.
- Once affix is added to the word it modifies the meaning accordingly.
- There are several types of affixes.
  1. Prefix
  2. Infix
  3. Suffix

AMEY THAKUR B-50 Amey.

Inflectional Morphology

- When an original word (stem) is fused with a grammatical morpheme the resultant word has a similar class. This is called an inflectional morpheme
- Affixes in inflectional are quite small.
- In english, words that get inflected are noun, verb and adjective
- Example!

①	Talk	Talks	Talking	Talked
②	Cry	Cries	Crying	Cried

Derivational Morphology

- When an original word is fused with an grammatical morpheme and the resultant has a whole new class. This is known as derivational morphology.
- Meaning of derivational morpheme is different from original word.
- In english, words that get derivated are nouns generated from adjectives
- Example.

	Adjective	Suffix	Noun
①	Kill	-er	Killer
②	Computerize	-ation	Computerization

AMEY THAKUR B-50 Amey

Q5. What is FST (Finite State Transducer)? What is their use in morphological analysis?

Ans:

Finite State Transducer

- FST is a type of Finite State Automata which has two set of symbols
- It is a two tape automation that recognizes or generates a pair of string as output, one from each string
- FST defines relationship between set of strings
- FST has two tapes on the table, i.e. input and output.
- FST is a multifunction device and can be viewed as

① Translator

- It reads one string on one tape and gives another string as output (tape).

② Generator

- It analyzes two tapes and gives as output a pair of strings as yes/no. according to how they match.

③ Recognizer

- It takes a pair of strings as two tapes and accepts or rejects according to their match.

④ Relator

- It compares the string available on two tapes

AMEY THAKUR B-50 Amey

- The two level morphology represents correspondence between lexical and surface levels.
- It has two tapes
  1. Input
  2. Output
- For morphological processing one tape holds lexical representation and another holds surface

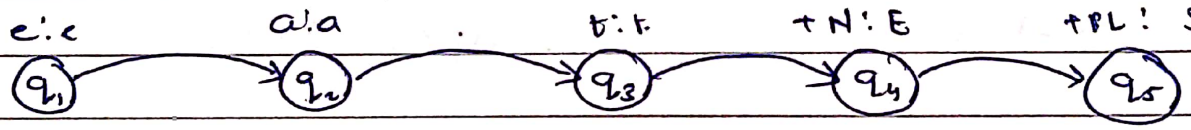
Lexical

C	A	T	+N	+PL	

Surface

C	A	T	S

- The transition of regular noun happens as follows:



- Through the morphological process the transformation table looks like

Input	Parsed Output
Cat	Cat + N + SG
Cats	Cat + N + PL
Geese	Goose + N + PL
Reading	Read + V + pres-part

AMEY THAKUR B-50 Amey

Q 6. What is stemming and lemmatization and explain their uses in NLP.

Ans:

Stemming

- It is a technique used to extract the base form of words by removing its affixes. It's like cutting a branch of tree from its stem.
- Example: Stem of words eaten, eating, eats is eat.
- The indiscriminate removal of affix from base morpheme can be successful sometimes and not always. that's why it has its limitations.
- Stemming is used in NLP in information retrieval system like search engine.
- It is used to determine domain vocabularies in domain analysis
- Stemming has various algorithms
  1. Porter's stemmer
  2. Dowson stemmer
  3. Lovins stemmer
  4. Krovetz stemmer

AMEY THAKUR B -50 Amey

Lemmatization

- This is a process of grouping together different forms of inflected words so they can be analyzed as a single item.
- Lemmatization is similar to stemming but it brings context to the words and links words with similar meanings
- It does morphological analysis of the words.
- Uses of Lemmatization in NLP are:
  1. In comprehensive retrieval system like search engine
  2. Used in compact indexing
- Examples:

  rocks - "rock"

  better - "good"

AMEY THAKUR B-50 Amey

Q7. What is RE (Regular Expression) ?
Explain the use of RE in morphological analysis.

Ans!

Regular Expression

- Also called as Regex.
- It is used for pattern matching standards for string parsing and replacement.
- Example! Used to parse email address, URLs, dates, logs, files, etc
- It contains an algebraic formula whose values is a pattern consisting of set of strings known as language of expression
- Simple type of RE contains single symbol or sequence of characters. Ex - /a/, /avengers/.

Brackets - Characters are group by putting them between square brackets. This way any character in the class will match one char in input.
Ex - /[abcd]/ will match any of a, b, c & d.

Range - Sometimes RE leads to cumbersome notation. It is used to specify range
Ex - /[abcdefghijklmnopqrstuvwxyz]/ - any lowercase letter.

- Regular expressions are case sensitive, the pattern /t/ matches t but not T. To solve this RE is mention as /[Tt]ree/.

- * or + - Use of * or + allows you to add one or more of preceeding character

AMEY THAKUR B-50 Amey

Anchor - These are special characters to accomplish string operations to start and end of a text

'^' - Specifies start of string

'$' - Specifies end of string

Special Characters

. → any character except one line

\w → any word character.

\d → any digit character

\D → Anything but a digit character

\b → A word boundary

\B → Anything but a word boundary

AMEY THAKUR B-50 Amey.

Q8. What is POS Tagging? Explain the different types of POS tagging

Ans:

POS Tagging

- Tagging is a kind of classification that may be defined as automatic assignment of description to the tokens.
- POS Tagging is simply a task of labelling each word in a sentence with its correct part of speech

Types of POS Tagging

① Rule based POS Tagging

- This type uses a dictionary or lexicon for getting possible tags for tagging each word
- As the name suggest, all kind of information in rule-based POS Tagging is coded in form of rules.

(a) Context Pattern rules

(b) As RE complied into FA, intersected with lexical representation

- It has two stage architecture

(a) First stage

- It uses a dictionary to assign each word a list of potential POS

(b) Second stage

- It uses large list of handwritten rules to sort down the list to single POS.

AMEY THAKUR B -50 Amey.

② Stochastic POS Tagging

- The model that includes frequency or statistics can be called as stochastic

ⓐ Word Frequency Approach

- The tag encountered the most frequently in the training set is assigned to the instance of the word.

ⓑ Tag Sequence Approach

- The best tag given for a word is determined by the probability that occurs in previous tags.

AMEY THAKUR B-50 Amey

Q9. What is N-Gram? How is it used for word disambigyation in NLP?

Ans:

N-Gram

- N-gram can be defined as contiguous sequence of n items from a given sample of text or speech.
- It predicts the probability of a given N-gram within any sequence of words in a language
- Example:

  Unigram - ("This", "article", "is", "on", "NLP")

  Bigram - ("This article", "article is", "is on", "on NLP")

- Now to establish a relationship on how to find a next word in a sentence, we need $p(w|h)$

  $P(\text{NLP} \mid \text{this article is on})$

- Generalizing ....

  $p(w_5 \mid w_1, w_2, w_3, w_4)$ or $p(w)$

  $= (P w_n \mid P w_1, w_2, \ldots w_n)$

- For calculation, we need chain rule of probability

  $P(A|B) = P(A,B) \mid P(B)$

- Generalizing...

  $P(w_1, w_2, w_3, \ldots w_n) = \pi_i P(w_i \mid w_1, w_2, \ldots w_n)$

- Using markov assumption

  $P(w_i \mid w_1, w_2, \ldots w_{i-1}) \approx P(w_i \mid w_{i-k} \ldots w_{i-1})$

→ For Unigram - $P(w_1, w_2, \ldots, w_n) \approx \pi_i P(w_i)$

→ For Bigram - $P(w_i \mid w_1, w_2, \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$