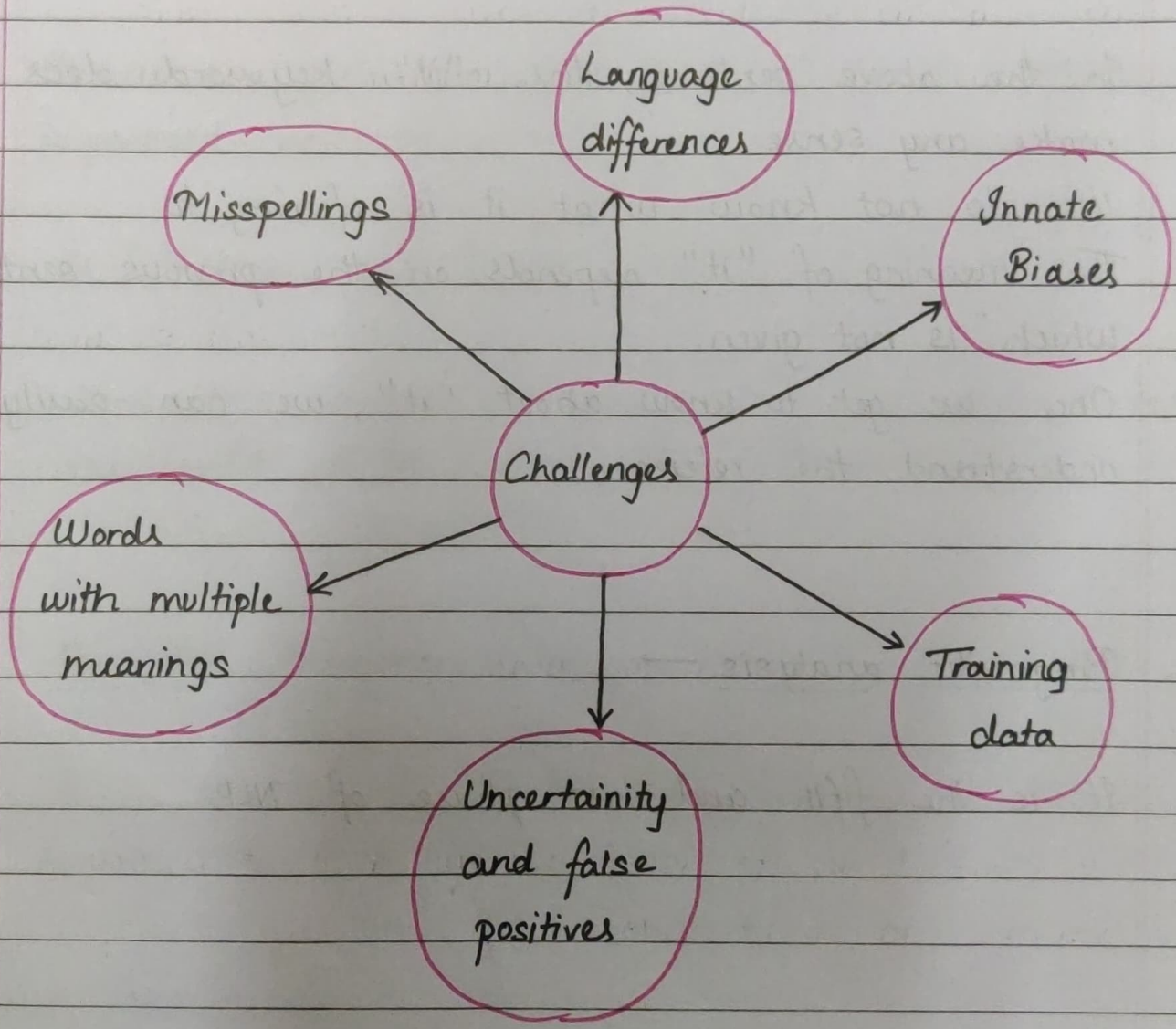
Q2. Challenges of NLP



★ Misspellings →

- Natural languges are full of
  - → misspellings
  - → typos
  - → inconsistencies in style

- The problem is compounded when you add accents or other characters that are not in your dictionary.

★ Language differences →

- An English speaker and an Italian speaker would say the same thing in different ways in their own languages

- Even though the two sentences have the same meaning, NLP wouldn't understand the latter unless you convert it into English first.

★ Words with multiple meanings →

- NLP is based on the assumption that language is precise and unambiguous.

- However in reality, language is neither precise, nor unambiguous.
- Many words have multiple meanings and can be used in different ways.
- For example, when ~~be~~ we say the word "bark", we could mean dog bark or tree bark.

★ Uncertainity and False Positives →

- False positives occur when the NLP detects a term that should be understandable, but cannot be replied to properly
- The goal is to create an NLP system that can identify it's limitations and clear up any confusion by using questions or hints.

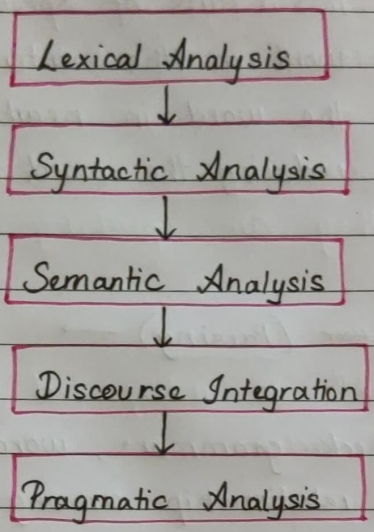
★ Training data →

- One of the biggest challenges with natural language processing is innaccurate training data.
- The more training data you have, the better your results will be
- If you give your system incorrect or biased data, it will learn from the wrong things or learn inefficiently.

Unit - 1

★ Long answers

Q1. Phases of NLP



1. Lexical Analysis (Morphological) →

- The first phase of NLP is Lexical Analysis
- This phase scans the source code as a stream of characters and converts them into meaningful lexemes
- It divides the whole text into paragraphs, sentences and words
- It involves identifying and analysing the structure of words

Lexicon of a language → collection of words and phrases in that particular language.

- We must perform Lexicon Normalisation
- The most common Lexicon Normalization techniques are →

★ Stemming → The process of reducing derived words to their word stem / base / root form.
(eg. derived word form written with "ing", "ly", "s' etc is reduced to it's root word)

★ Lemmatization → Process of reducing a group of words into their dictionary form or lemma

→ It takes into account the things like POS (Parts of Speech), the meaning of the word in the sentence, the meaning of the word in nearby sentences, etc. before reducing the word to it's lemma.

2. Syntactic Analysis (Parsing) →

- It is used to check grammar, word arrangements, and it shows the relationship among the words.
- Syntactical parsing involves analysing the words in a sentence for grammar.
- Dependency grammar and parts of speech (POS) tags are important attributes

eg. "Chennai goes to the John"

Since this sentence does not make any sense in the real world, it is rejected by syntactic analyzer.

3. Semantic Analysis →

- It is the process of ~~recieving~~ retrieving meaningful insights from text

- It mainly focusses on the literal meaning of words, phrases and sentences.
- It retrieves the possible meanings of a sentence that is clear and semantically correct.

4. Discourse Integration →

- Discourse Integration depends on the meanings of the sentences that preceeds it as well as those that follow it
- It is nothing but a sense of context

eg. "Ram wants it"

In the above sentence, the "it" keyword does not make any sense.
We do not know what it is referring to
The meaning of "it" depends on the previous sentence which is not given.
Once we get to know about "it", we can easily understand the reference.

5. Pragmatic analysis →

- It is the fifth and last phase of NLP

- It helps discover the intended effect by applying a set of rules that characterize dialogues.

- It is the study of meanings in a language and the extraction of insights from the text

- It understands how people communicate with eachother, in which context they are speaking and many other factors.

eg. "open the door" is interpreted as a request instead of an order.

Grammar

Grammar is defined as the rules for forming well structured sentences.

Grammar in NLP is a set of rules for constructing sentences in a language used to understand and analyze the structure of the sentences in text data.

Grammar also plays an essential role in describing the syntactic structure of well-formed programs like denoting the synathic rules used for conversation in natural language.

This includs identifying Parts of speech such as nouns, pronoun, verbs and Adjectives, determining the subjects and predicate the sentence, and identifying the relationships b/w words and phrases.

Mathematical, a grammar G can written as a 4-tuple (N, T, S, P) where:

=> N : set of non-terminal symbols or variables

=> T : set of terminal symbols

=> S : start symbol where S ∈ N

=> P : Production rules for Terminals as well as non-terminals.

=> It has in the form of $\alpha \rightarrow \beta$, where α and β are strings on $N \cup T$, and at atleast one symbol of α belongs to N

Types of Grammar in NLP

=> Context-free Grammar

=> Constituency Grammar

=> Dependency Grammar

Context-free Grammar:-

Context free Grammars (CFGs) are formal a system used in NLP to describing both the syntactic structure of human language and aprogramming language.

Set of rules

=> A CFG's Consists of some set of written rules

$A \rightarrow B_1, B_2 \ldots B_n$ where symbols appear on the left side hand side (LHS) are called "non-terminals" and the symbols on the right hand side (RHS) can be either non-terminals or terminals.

=> The terminals symbols are those that don't appear on the left-hand side of any rule, because they are "atomic" for the language.

=> In a grammar, for a human language the terminal are the words and the non-terminals are phrase which are (NP, VP and S

Ex:-

A Parse of the sentence "the graffic dreams"

Grammar

S → NP VP

NP → DET X

VP → Z NP

VP → iv

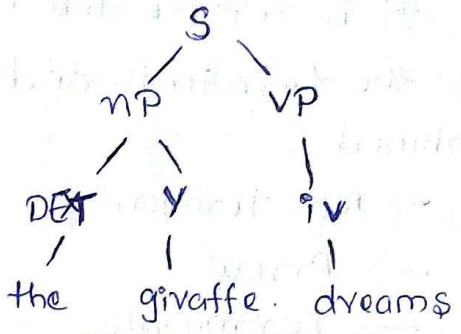
det → the
→ a
→ an

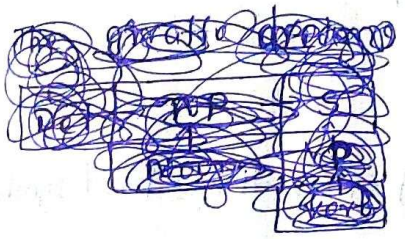
Y → giraffe
→ apple

iv → dreams

tv → eats
→ dreams.

Parse tree





S ⇒ NP VP
⇒ X Y VP
⇒ The Y VP
⇒ The giraffe VP
⇒ The giraffe iv
⇒ The giraffe dreams

* We can see that the root of every subtree has a grammatical category that appears on the left-hand side of a rule, and.
* the children of that root are identical to the elements on the right-hand side of that rule.

CFG consists of a finite set of grammar rules having the following four components.

⇒ Set of non-terminals :

It represents by V. The non-terminals are syntatic variables that denote the sets of strings, which help define the language generated with the help of grammar.

⇒ Set of Terminals :-

It is also known as Tokens, and it is represented by Σ. Strings are formed with the help of the basic symbols of terminals.

=> Set of Production

It is represented by P. The set explains the how the terminals and the non-terminals can be combined.

-> Non-terminals
-> Arrow
-> Terminals
-> LHS called non-terminals and RHS Production called terminal

=> Start Symbol:-

The Production begins from the start symbol. It is represented by symbols.

Non-terminals symbols are always designated as start symbols.

Constituency Parsing

Constituency Parsing is a natural language Processing technique that is used to analyze the grammatical structure of sentences.

It is a type of syntactic Parsing that aims to identify the constituents, or subparts, of a sentence and the relationships b/w them.

The output of a Constituency Parser is typically a Parse tree, which represents the hierarchical structure of the sentence.

The Process of Constituency Parsing involves identifying the syntactic structure of a sentence by analyzing its words and phrases.

=> Typically involves identifying the noun Phrases, verb Phrases and other constituents, and then determining the relationships b/w them.

Constituency Parsing is an important step in natural language Processing and is used in a coide range of applications such as natural language under -standing, machine translation, and text Summariza -tion.

Ex:- The cat discovered a fish"

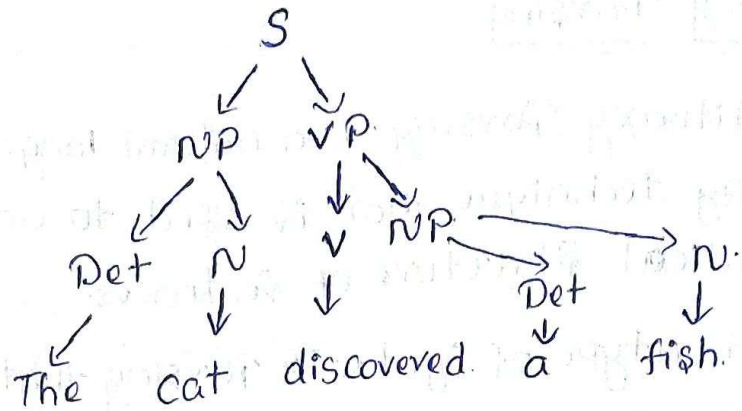
Nouns - Cat, fish
Verb - discovered
Det : The, a

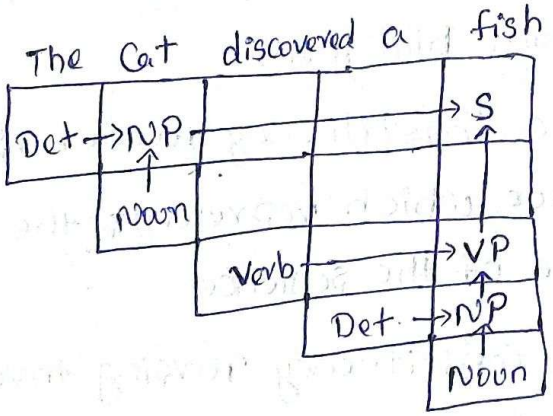
Phrase structure

S → NP VP

NP → Det N

VP → V NP





from this method we will Construct this Phrase structure.

⇒ When Det value and Noun are occur then it is Nounphrase.

⇒ when Nounphrase and verb are occur then it is verb phrase.

⇒ when Nounphrase and verbphase are occur then it is an "S".

Dependency grammar:-

Dependency Parsing is a natural language Processing technique that is used to analyze grammatical structure of sentences.

It is a type of syntactic Parsing that aims to identify the relationships, or dependencies, b/w words in a sentence.

The output of a dendency Parser, is typically a dependency tree or a graph, which represents the linear structure of the sentence.

This typically involves identifying the subject, object and other grammatical elements and then determining the relationships blw them.

Ex:- Fruit flies like a banana.