12-10-2021

- Grammars & derivation
- Parsing ← Robust Parsing
  Shallow parsing
  Probabilistic parsing

Grammar → $G = (V, \Sigma, S, P)$, $P = \{ S \to aSb, S \to \varepsilon \}$

Is $aabb \in L(G)$? $\Sigma = \{a, b\}$, $V = \{S\}$

$S \overset{(1)}{\Rightarrow} aSb \overset{(2)}{\Rightarrow} aaSbb \overset{(3)}{\Rightarrow} aabb \in L(G)$.

Grammar is for generating the sentences. But, $abab \notin L(G)$.
or used for deriving the sentence.

Syntax ≡ Format or structure.

Whether the sentence is correct as per syntax, we do it using grammar. This process is called syntax analysis.

In the above (1), (2), (3) are configurations of derivation process.
The PDA if used to recognize, then also it moves through

sequence of configurations; if input is accepted, then the last configuration is accepting configuration.

The other part of Grammar is semantic analysis, i.e. what is meaning associated with that sentence.

Grammars

Regular Grammar	— type – 3	– FA
context-free Grammar	– 2	– PDA
Context-sensitive or	– 1	– LBA
Unrestricted Grammar,	– 0	– TM

$A \in V$.

Type-3

$A \to aA$, $\therefore w = aa$, $w = a$, $w = aaaa$.

$A \to \varepsilon$

Type-2

$A \to (V \cup \Sigma)^*$

Type-1

$(V \cup \Sigma)^* \to (V \cup \Sigma)^*$, with at least one Var in LHS, $|\alpha| \leq |\beta|$

(α under the first $(V \cup \Sigma)^*$, β under the second)

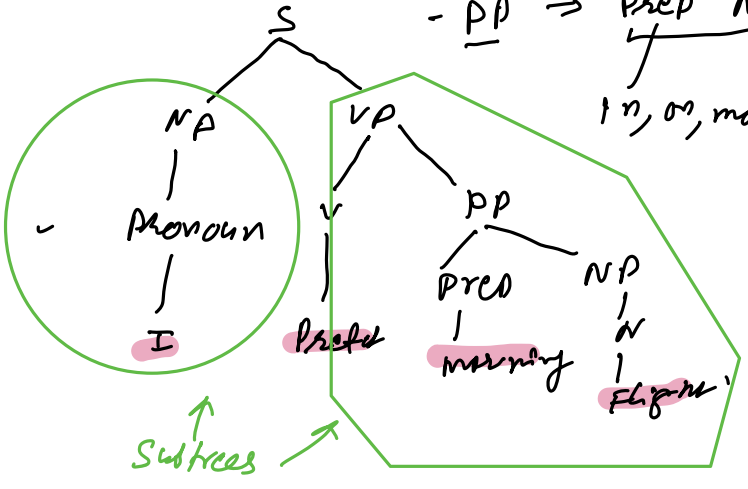
type-0 : $\alpha \to \beta$, with no restriction like $|\alpha| \leq |\beta|$.

But in NLP, only CFG is used, because the type 0 and 1 grammars are very difficult to implement. But, in fact, Nat. Lang. is Context-sensitive (type - 1).

CFG: $S \to NP\ VP$; I prefer morning flight

- $VP \to V\ NP$; Prefer a morning flight.
- $VP \to V\ NP\ PP$; Leaves Mumbai in the morning
- $VP \to V\ PP$ (V, NP, Prep NP = PP)
- $PP \to Prep\ NP$ → Leaving on Monday (Prep NP = PP)

in, on, morning



- If wrong rule has been used, then we backtrack.
- If nothing works, then fail.

A Good parsing must :-

- Construct parse tree, its subtrees, and further subtrees.

↑ "Chunking", i.e. segmentation & tagging

- It must do disambiguations

When parsing is correctly done, like above, it is called "Robust parsing".

Challenge → As the sentence becomes larger & larger, it is highly time consuming to parse it. Hence, it cannot be used in real time speech and in machine translations.

In that case we do not go for complete parsing, but Partial, we break the sentence and parse its units.

"The book is on table, it will remain there if you ↑ 3-4 more sentences don't collect it"

This is called "Shallow parsing", this uses large corpus of text where millions of small size parsed sentences are already available. Thus, a parser need to search those already parsed sentences. (Corpora)

It is annotated.

Present trend is shallow & probabilistic parsing.

Corpus : NLTK. (3-4 GB)

annotated: Book [n] is [Aux] on [Prep] table [N]
Example.

—— x —— x ——