Syntactic and Statistical parsing

Syntactic & statistical Parsing

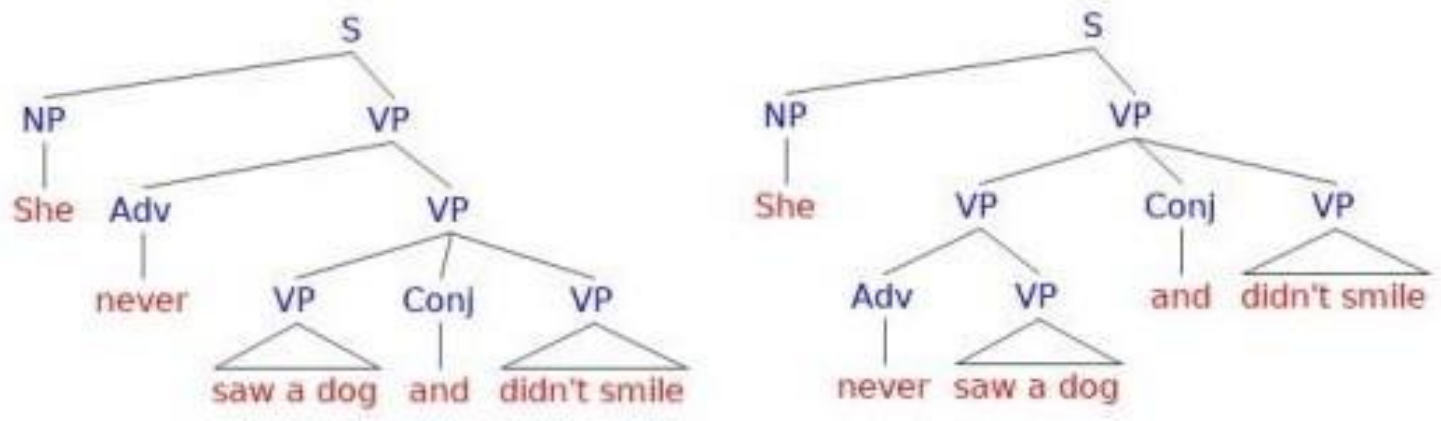
- ➢ What is Syntactic parsing,
- ➢ Ambiguity,
  - Attachment Ambiguity,
  - Coordination Ambiguity.
- ➢ CKY Parsing
  - Conversion to Chomsky Normal Form,
- ➢ Probabilistic Context-Free Grammars,
- ➢ Probabilistic CFG Parsing
  - Ambiguity in CFG
  - Managing ambiguity in CFG

Syntactic Parsing (Constituency)

- It refers to the breaking down of a text or sentences into its constituents.
- Composed of:
- Terminals(words)
- Non-terminals(phrases/sentences)
- Parse trees are used in grammar checking while word processing.

Ambiguity

- Structural ambiguity : It is similar to the grammatical phrase structuring.
- It occurs when a sentence has a grammar that can be parsed more than once for e.g.



- The two types of ambiguity are:

1. **Attachment ambiguity:** Constituent or a part of a sentence that can be attached to a parse tree multiple times.
2. **Coordination ambiguity:** Conjunctions are used to join two different phrases. e.g. the thief shot the jeweler **and** the cop panicked.

S → NP VP

S → Aux NP VP

S → VP

NP → Pronoun

NP → Proper-Noun

NP → Det Nominal

Nominal → Noun

Nominal → Nominal Noun

Nominal → Nominal PP

VP → Verb

VP → Verb NP

VP → Verb NP PP

VP → Verb PP

VP → VP PP

PP → Preposition NP

Det → that | this | a

Noun → book | flight | meal | money

Verb → book | include | prefer

Pronoun → I | she | me

Proper-Noun → Houston | TWA

Aux → does

Preposition → from | to | on | near | through

CKY Parsing or Cocke-Kasami-Younger algorithm

Handles syntactic disambiguation with the help of dynamic programming.

Chomsky normal form

- At the beginning of the CKY algorithm we need to convert the CFGs to CNF.
- Unit productions are formed when there is a single non terminal towards the right.

$$S \rightarrow Aux\ NP\ VP$$

$$S \rightarrow XI\ VP$$

$$XI \rightarrow Aux\ NP$$

$\mathscr{L}_1$ Grammar	$\mathscr{L}_1$ in CNF
$S \rightarrow NP\ VP$	$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$	$S \rightarrow X1\ VP$
	$X1 \rightarrow Aux\ NP$
$S \rightarrow VP$	$S \rightarrow book \mid include \mid prefer$
	$S \rightarrow Verb\ NP$
	$S \rightarrow X2\ PP$
	$S \rightarrow Verb\ PP$
	$S \rightarrow VP\ PP$
$NP \rightarrow Pronoun$	$NP \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det\ Nominal$	$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$	$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal\ Noun$	$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$	$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$	$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb\ NP$	$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$	$VP \rightarrow X2\ PP$
	$X2 \rightarrow Verb\ NP$
$VP \rightarrow Verb\ PP$	$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$	$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$	$PP \rightarrow Preposition\ NP$

1. Copying all the conforming rules to a new grammar.
2. Conversion of terminals to non-terminals.
3. Conversion of unit products.
4. All the new rules are made binary and added to new grammar.

Statistical Constituency Parsing

- It is possible to build probabilistic parsers consisting of syntatic knowledge.
- PFGCs or probabilistic context-free grammar which is an enhancement of CFG.
- Aprobability is assigned to each rule.
- These are trained on treebank grammars.
- Non-terminals are made more specific or more general.
- Also known as Stochastic context-free grammar SCFG.
- Consists of N= non-terminal symbols, Σ=terminal symbols, R=rules or productions $A \rightarrow \beta[p]$, β=string of symbols $(\Sigma \cup N)^*$, and p is a number between 0 and 1 $P(\beta|A)$ and S= start symbol.
- R is augmented with a conditional probability:

$$A \rightarrow \beta[p]$$
$$P \rightarrow (A \rightarrow \beta) \qquad \Sigma P(A \text{ - > } \beta)$$
$$P \rightarrow (A \rightarrow \beta|A)$$
$$P(RHS|LHS)$$

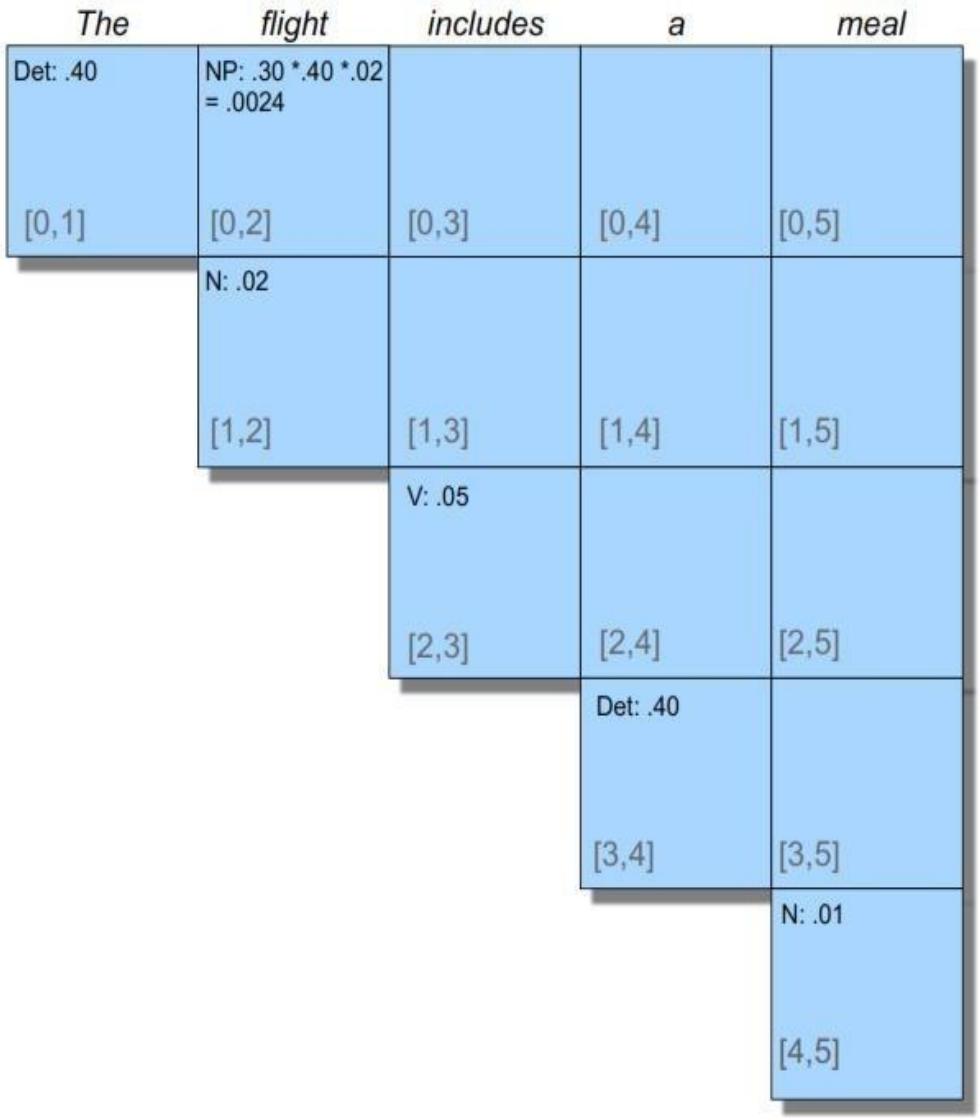
Probabilistic CKY Parsing PCFGs

- The probabilistic CKY concludes that the PCFG is in the Chomsky normal form.
- Indices are assumed between each word.

⓪ Book ① the ② flight ③ through ④ Houston ⑤

- These are considered and each constituent in the CKY parse tree is encoded in a two dimensional matrix.
- The upper triangular portion of the matrix is used (n+1)x(n+1) matrix.
- Each cell table [i,j] contains a list of constituents that spans a sequence of words from i to j.
- The sentence " the flight includes a meal" is connected to chomsky normal form in order for the CKY algorithm to work on it and to handle the rule properties.
- Separate counts are needed for each constituents in the PCGs using the Inside-out algorithm.

Probabilistic CKY Parsing PCFGs



S	$\rightarrow$	$NP\ VP$	.80	Det	$\rightarrow$	*the*	.40
NP	$\rightarrow$	$Det\ N$	.30	Det	$\rightarrow$	*a*	.40
VP	$\rightarrow$	$V\ NP$	.20	N	$\rightarrow$	*meal*	.01
V	$\rightarrow$	*includes*	.05	N	$\rightarrow$	*flight*	.02

Problems with PCFGs

- **Poor independence assumptions:** CFG rules impose an independence assumption on probabilities that leads to poor modeling of structural dependencies across the parse tree.

- **Lack of lexical conditioning**: CFG rules don't model syntactic facts about specific words, leading to problems with sub categorization ambiguities, preposition attachment, and coordinate structure ambiguities.

Probabilistic CFG Parsing

- Lexicalized grammar frameworks such as CFG pose problems for which the phrase based methods we've been discussing are not particularly well-suited.

$$X/Y \; Y \Rightarrow X$$

$$Y \; X\backslash Y \Rightarrow X$$

Ambiguity in CFG