# SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 98
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

## Department of Artificial Intelligence and Data Science

### Course Name – 19AD601 – Natural Language Processing

### III Year / VI Semester

### Unit 3 – SYNTACTIC ANALYSIS

### Topic 8- Probabilistic Lexicalized CFGs

# Probabilistic Lexicalized CFGs

Probabilistic Lexicalized CFGs

- PCFGs turn out to be a rather poor model for statistical parsing. Lexicalized PCFGs, which build directly on ideas from regular PCFGs, but give much higher parsing accuracy.

- Weaknesses of PCFGs:
- 1) lack of sensitivity to lexical information; and
- 2), lack of sensitivity to structural preferences.

- The basic idea in lexicalized PCFGs will be to replace rules such as
- S → NP VP
- with lexicalized rules such as

# Probabilistic Lexicalized CFGs

**If** the rule contains NN, NNS, or NNP:
Choose the rightmost NN, NNS, or NNP

**Else If** the rule contains an NP: Choose the leftmost NP

**Else If** the rule contains a JJ: Choose the rightmost JJ

**Else If** the rule contains a CD: Choose the rightmost CD

**Else** Choose the rightmost child

Figure 6: Example of a set of rules that identifies the head of any rule whose left-hand-side is an NP.

# Probabilistic Lexicalized CFGs

**If** the rule contains Vi or Vt: Choose the leftmost Vi or Vt

**Else If** the rule contains a VP: Choose the leftmost VP

**Else** Choose the leftmost child

Figure 7: Example of a set of rules that identifies the head of any rule whose left-hand-side is a VP.

# Probabilistic Lexicalized CFGs

Thus we have replaced simple non-terminals such as S or NP with lexicalized non- terminals such as S(examined) or NP(lawyer).

Each rule in the lexicalized PCFG will have an associated parameter, for ex-

ample the above rule would have the parameter

q(S(examined) → NP(lawyer) VP(examined))

There are a very large number of parameters in the model, and we will have to take some care in estimating them: the next section describes parameter estimation methods.

Each rule in the lexicalized PCFG has a non-terminal with a head word on the left hand side of the rule: for example the rule

S(examined) → NP(lawyer) VP(examined)

has S(examined) on the left hand side.

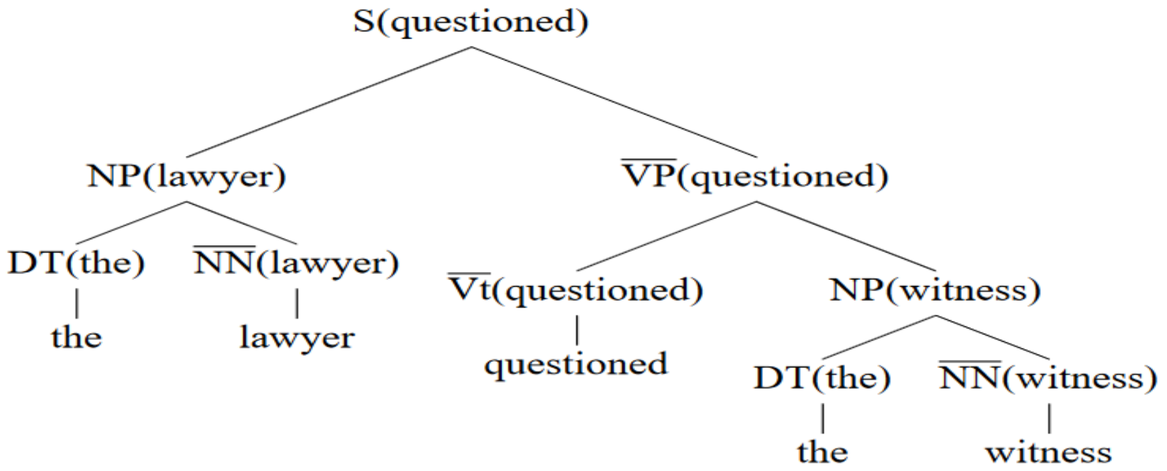
# Probabilistic Lexicalized CFGs

A lexicalized PCFG in Chomsky normal form is a 6-tuple G = (N, Σ, R, S, q, γ) where:

- N is a finite set of non-terminals in the grammar.
- Σ is a finite set of lexical items in the grammar.
- R is a set of rules. Each rule takes one of the following three forms:

1. X(h) →1 Y1(h) Y2(m) where X, Y1, Y2 ∈ N , h, m ∈ Σ.
2. X(h) →2 Y1(m) Y2(h) where X, Y1, Y2 ∈ N , h, m ∈ Σ.
3. X(h) → h where X ∈ N , h ∈ Σ.

# Probabilistic Lexicalized CFGs




In this case the parse tree consists of the following sequence of rules:

$S(questioned) \rightarrow_2 NP(lawyer)\ VP(questioned)$
$NP(lawyer) \rightarrow_2 DT(the)\ NN(lawyer)$
$DT(the) \rightarrow the$
$NN(lawyer) \rightarrow lawyer$
$VP(questioned) \rightarrow_1 Vt(questioned)\ NP(witness)$
$NP(witness) \rightarrow_2 DT(the)\ NN(witness)$
$DT(the) \rightarrow the$
$NN(witness) \rightarrow witness$

# THANK YOU