CS344: Introduction to Artificial Intelligence

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 30: Probabilistic Parsing: Algorithmics

(Lecture 28-29: two hours on student seminars on Default Reasoning, Child Language Acquisition and Short Term and Long Term Memory)

Formal Definition of PCFG

- A PCFG consists of
  - A set of terminals $\{w_k\}$, k = 1,....,V

    $\{w_k\}$ = { child, teddy, bear, played...}
  - A set of non-terminals $\{N^i\}$, i = 1,...,n

    $\{N_i\}$ = { NP, VP, DT...}
  - A designated start symbol N^1
  - A set of rules $\{N^i \rightarrow \zeta^j\}$, where ζ^j is a sequence of terminals & non-terminals

    NP $\rightarrow$ DT NN
  - A corresponding set of rule probabilities

Rule Probabilities

- Rule probabilities are such that

$$\forall i \sum_{i} \mathrm{P}(\mathrm{N}^{i} \rightarrow \zeta^{j}) = 1$$

E.g., P(NP → DT NN) = 0.2
P(NP → NN) = 0.5
P(NP → NP PP) = 0.3

- P(NP → DT NN) = 0.2
  - Means 20 % of the training data parses use the rule NP → DT NN

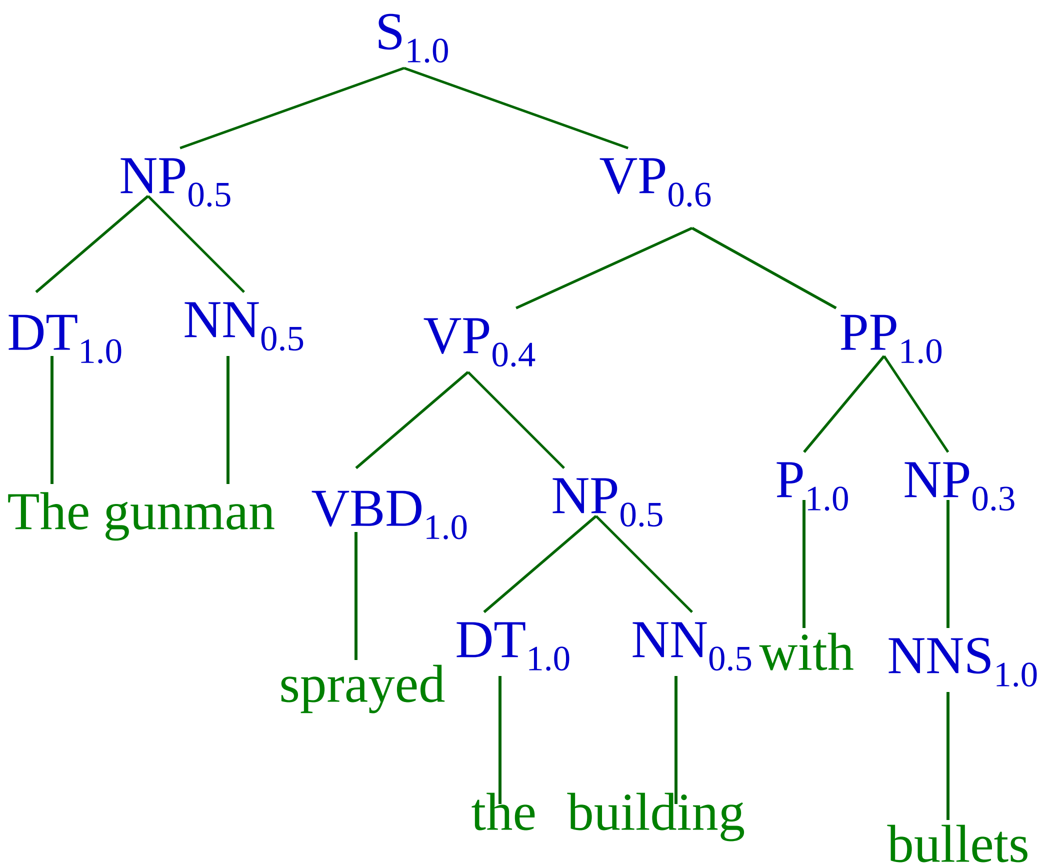
Probabilistic Context Free Grammars

- $S \rightarrow NP\ VP$ 1.0
- $NP \rightarrow DT\ NN$ 0.5
- $NP \rightarrow NNS$ 0.3
- $NP \rightarrow NP\ PP$ 0.2
- $PP \rightarrow P\ NP$ 1.0
- $VP \rightarrow VP\ PP$ 0.6
- $VP \rightarrow VBD\ NP$ 0.4

- $DT \rightarrow$ the 1.0
- $NN \rightarrow$ gunman 0.5
- $NN \rightarrow$ building 0.5
- $VBD \rightarrow$ sprayed 1.0
- $NNS \rightarrow$ bullets 1.0

Example Parse t_1`

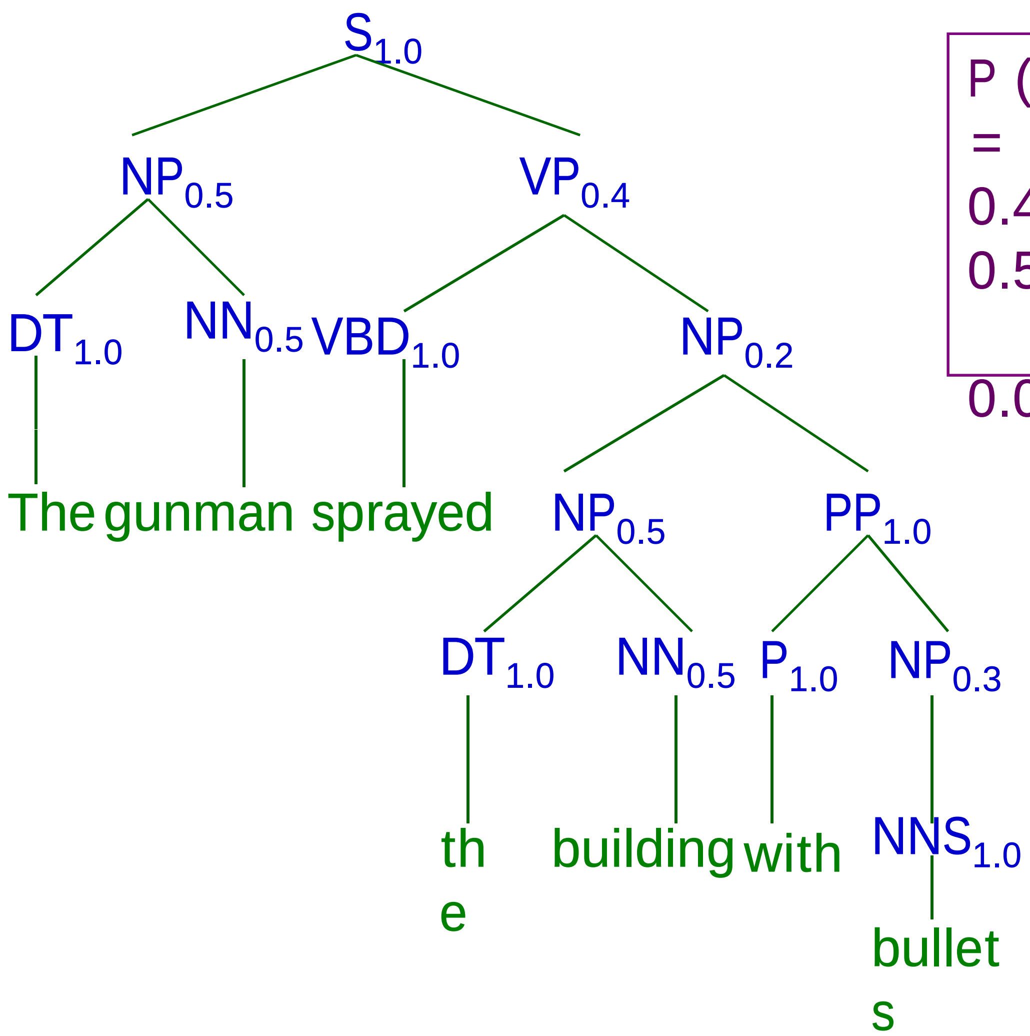
- The gunman sprayed the building with bullets.



$P(t_1)$ = 1.0 * 0.5 * 1.0 * 0.5 * 0.6 * 0.4 * 1.0 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0 = 0.00225

Another Parse t_2

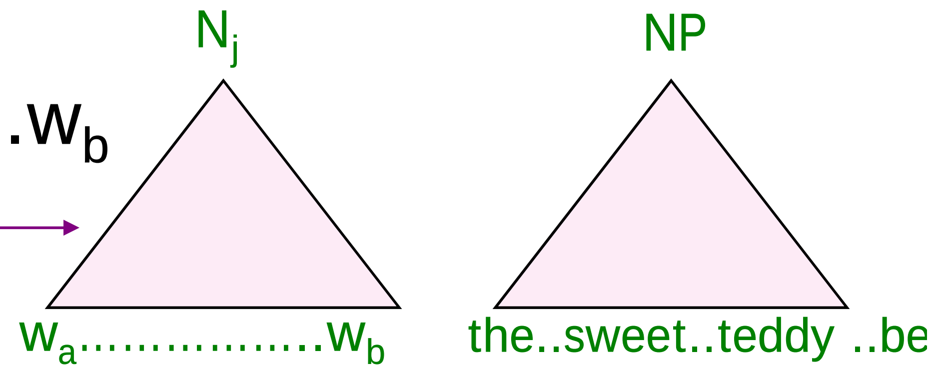
- The gunman sprayed the building with bullets.



$P(t_2)$
= 1.0 * 0.5 * 1.0 * 0.5 * 0.4 * 1.0 * 0.2 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0
= 0.0015

Probability of a sentence

- Notation :
  - w_{ab} – subsequence $w_a \ldots w_b$
  - N_j dominates $w_a \ldots w_b$ or $yield(N_j) = w_a \ldots w_b$



- Probability of a sentence = $P(w_{1m})$

$$P(w_{1m}) = \sum_t P(w_{1m}, t)$$

Where t is a parse tree of the sentence

$$= \sum_t P(t) P(w_{1m} \mid t)$$

$$= \sum_{t: yield(t) = w_{1m}} P(t) \qquad \because P(w_{1m} \mid t) = 1$$

If t is a parse tree for the sentence w_{1m}, this will be 1 !!

Assumptions of the PCFG model

- Place invariance :

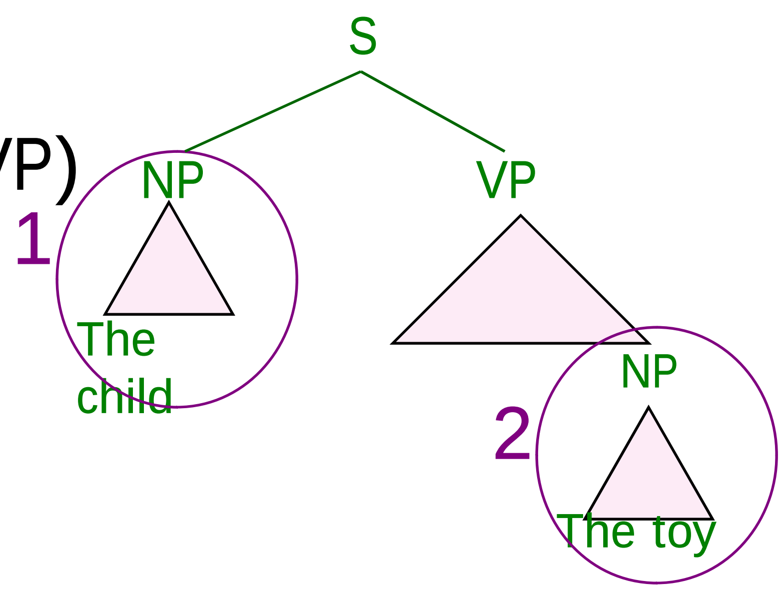
  P(NP → DT NN) is same in locations 1 and 2

- Context-free :

  P(NP → DT NN | anything outside “The child”)
  = P(NP → DT NN)

- Ancestor free : At 2,

  P(NP → DT NN|its ancestor is VP)
  = P(NP →DT NN)

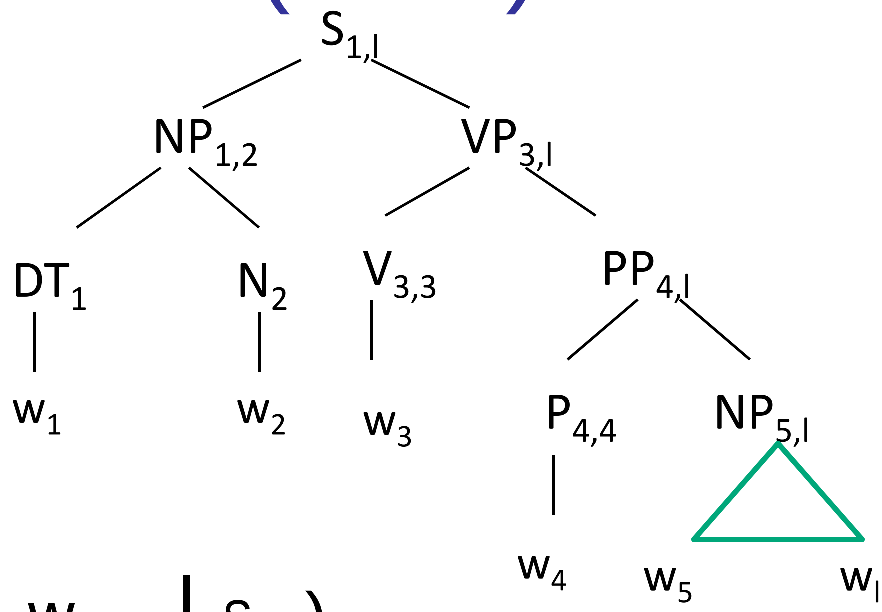


Probability of a parse tree

- *Domination* :We say N_j dominates from k to l, symbolized as $N^j_{k,l}$, if $W_{k,l}$ is derived from N_j
- P (tree |sentence) = P (tree | $S_{1,l}$)

  where $S_{1,l}$ means that the start symbol S dominates the word sequence $W_{1,l}$
- P (t |s) approximately equals joint probability of constituent non-terminals dominating the sentence fragments (next slide)

Probability of a parse tree (cont.)



$P (t|s) = P (t \mid S_{1,l})$

$= P (NP_{1,2}, DT_{1,1} , w_{1,}$
$N_{2,2}, w_{2,}$
$VP_{3,l}, V_{3,3} , w_{3,}$
$PP_{4,l}, P_{4,4} , w_{4,}\ NP_{5,l}, w_{5...l} \mid S_{1,l})$

$= P (NP_{1,2} , VP_{3,l} \mid S_{1,l}) * P (DT_{1,1} , N_{2,2} \mid NP_{1,2}) * D(w_1 \mid DT_{1,1}) * P (w_2 \mid N_{2,2}) * P (V_{3,3}, PP_{4,l} \mid VP_{3,l}) * P(w_3 \mid V_{3,3}) * P(P_{4,4}, NP_{5,l} \mid PP_{4,l}) * P(w_4|P_{4,4}) * P (w_{5...l} \mid NP_{5,l})$

(Using Chain Rule, Context Freeness and Ancestor Freeness)

HMM $\leftrightarrow$ PCFG

- O observed sequence $\leftrightarrow$ w_{1m} sentence
- X state sequence $\leftrightarrow$ t parse tree
- μ model $\leftrightarrow$ G grammar

- Three fundamental questions

HMM ↔ PCFG

- How likely is a certain observation given the model? ↔ How likely is a sentence given the grammar?

$$P(O \mid \mu) \leftrightarrow P(w_{1m} \mid G)$$

- How to choose a state sequence which best explains the observations? ↔ How to choose a parse which best supports the sentence?

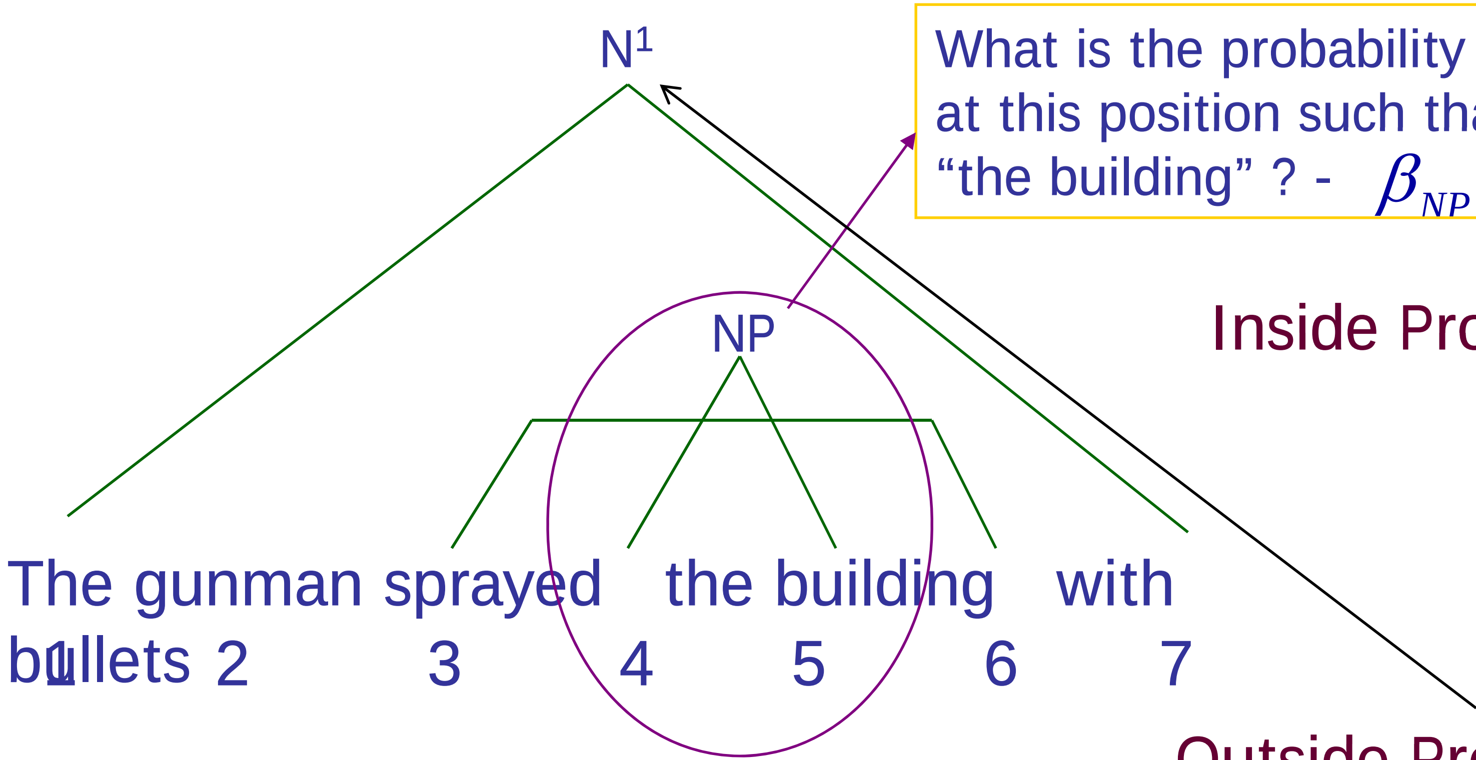
$$\arg\max_X P(X \mid O, \mu) \leftrightarrow \arg\max_t P(t \mid w_{1m}, G)$$

HMM $\leftrightarrow$ PCFG

- How to choose the model parameters that best explain the observed data? $\leftrightarrow$ How to choose rule probabilities which maximize the probabilities of the observed sentences?

$$\arg\max_{\mu} P(O \mid \mu) \leftrightarrow \arg\max_{G} P(w_{1m} \mid G)$$

Interesting Probabilities



What is the probability of having a NP at this position such that it will derive "the building" ? - $\beta_{NP}(4,5)$

Inside Probabilities

Outside Probabilities

What is the probability of starting from N^1 and deriving "The gunman sprayed", a NP and "with bullets" ? - $\alpha_{NP}(4,5)$

Interesting Probabilities

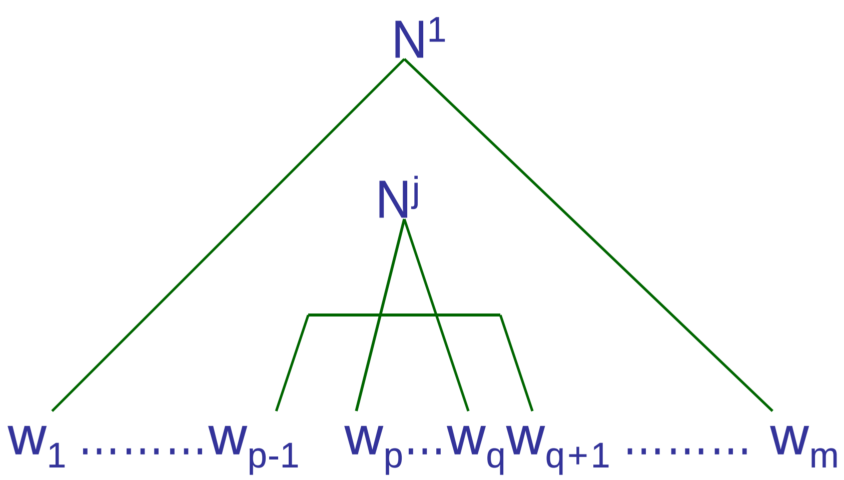
- Random variables to be considered
  - The non-terminal being expanded.
    E.g., NP
  - The word-span covered by the non-terminal.
    E.g., *(*4,5) refers to words "the building"

- While calculating probabilities, consider:
  - The rule to be used for expansion :
    E.g., NP $\rightarrow$ DT NN
  - The probabilities associated with the RHS non-terminals : *E.g.*, DT subtree's inside/outside probabilities & NN subtree's inside/outside probabilities

Outside Probability

- $\alpha_j(p,q)$: The probability of beginning with N^1 & generating the non-terminal N^j_{pq} and all words outside $w_p..w_q$

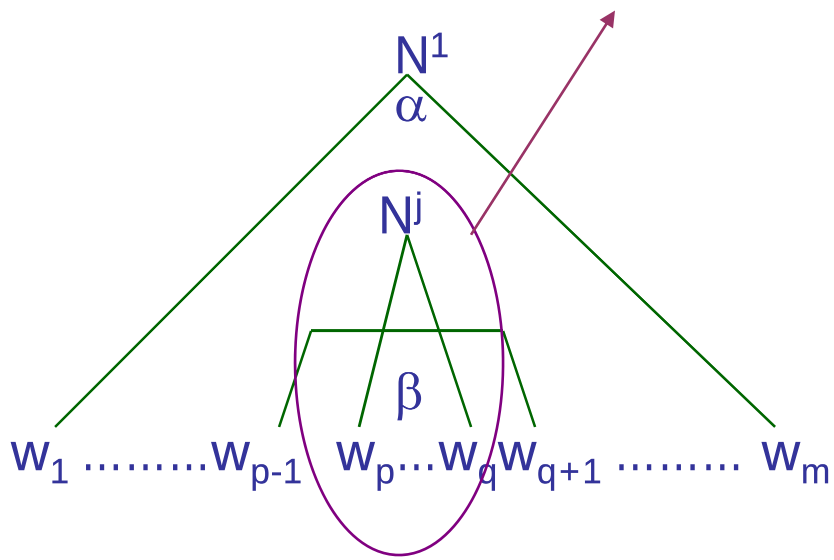
$$\alpha_j(p,q) = P(w_{1(p-1)}, N^j_{pq}, w_{(q+1)m} \mid G)$$



Inside Probabilities

- $\beta_j(p,q)$: The probability of generating the words $w_p..w_q$ starting with the non-terminal N^j_{pq}.

$$\beta_j(p,q) = P(w_{pq} \mid N^j_{pq}, G)$$

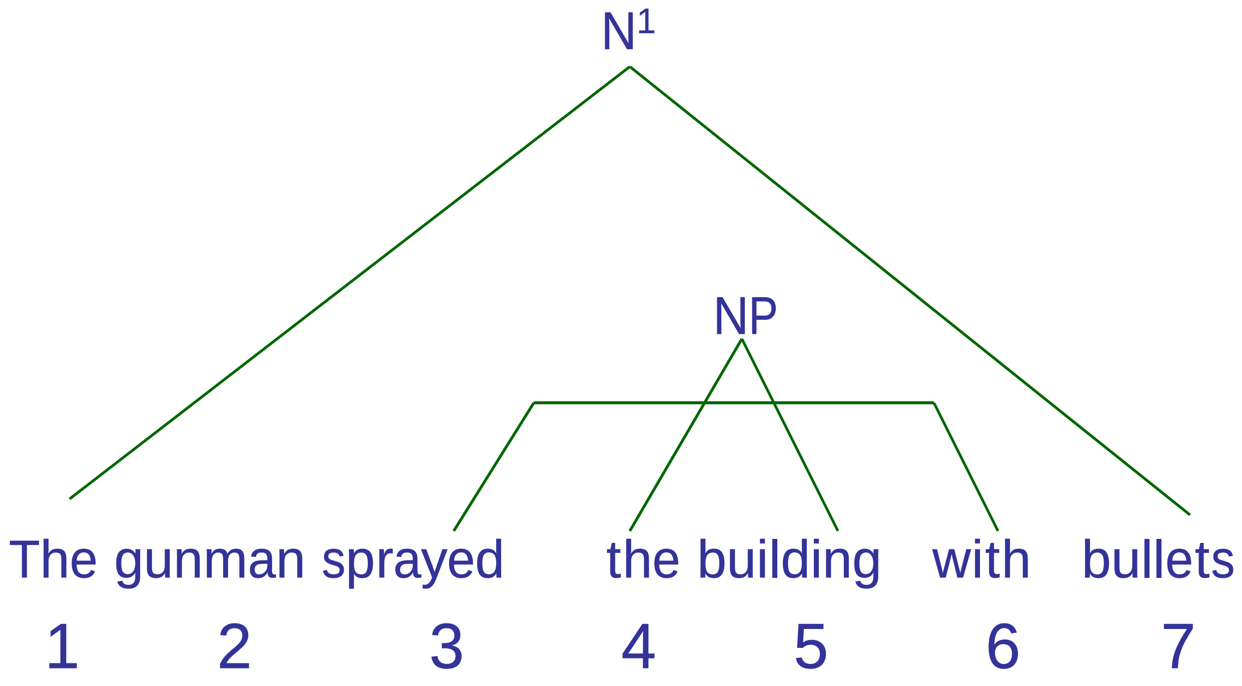


Outside & Inside Probabilities:example

$$\alpha_{NP}(4,5) \text{ for "the building"}$$

$$= P(\text{The gunman sprayed}, NP_{4,5}, \text{with bullets} \mid G)$$

$$\beta_{NP}(4,5) \text{ for "the building"} = P(\text{the building} \mid NP_{4,5}, G)$$



Inside probabilities $\beta_j(p,q)$

Base case:

$$\beta_j(k,k) = P(w_k \mid N^j_{kk}, G) = P(N^j_{kk} \rightarrow w_k \mid G)$$

- Base case is used for rules which derive the words or terminals directly

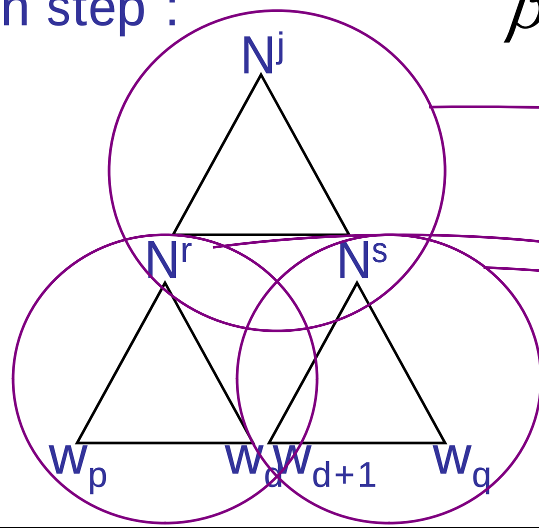
E.g., Suppose N^j = NN is being considered & NN $\rightarrow$ building is one of the rules with probability 0.5

$$\beta_{NN}(5,5) = P(building \mid NN_{5,5}, G)$$

$$= P(NN_{5,5} \rightarrow building \mid G) = 0.5$$

Induction Step

Induction step :



$$\beta_j(p,q) = P(w_{pq} \mid N^j_{pq}, G)$$

$$= \sum_{r,s} \sum_{d=p}^{q-1} P(N^j \rightarrow N^r N^s) * \beta_r(p,d) * \beta_s(d+1,q)$$

- Consider different splits of the words - indicated by *d*

  E.g., the huge building

  Split here for d=2 d=3

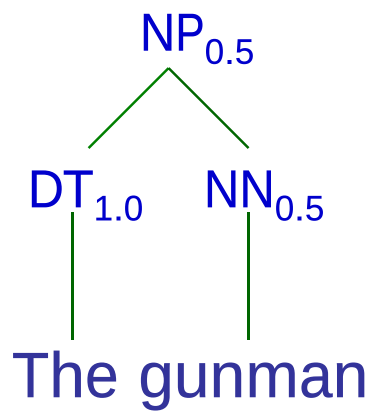
- Consider different non-terminals to be used in the rule:

  NP → DT NN, NP → DT NNS are available options

  Consider summation over all these.

The Bottom-Up Approach

- The idea of induction
- Consider “the gunman”



- Base cases : Apply unary rules

  DT $\rightarrow$ the Prob = 1.0

  NN $\rightarrow$ gunman Prob = 0.5

- Induction : Prob that a NP covers these 2 words

  = P (NP $\rightarrow$ DT NN) * P (DT deriving the word “the”) * P (NN deriving the word “gunman”)

  = 0.5 * 1.0 * 0.5 = 0.25

Parse Triangle

- A parse triangle is constructed for calculating $\beta_j(p,q)$
- Probability of a sentence using $\beta_j(p,q)$:

$$P(w_{1m} \mid G) = P(N^1 \rightarrow w_{1m} \mid G) = P(w_{1m} \mid N^1_{1m}, G) = \beta_1(1,m)$$

Parse Triangle

	The (1)	gunman (2)	sprayed (3)	the (4)	building (5)	with (6)	bullets (7)
1	$\beta_{DT}=1.0$						
2		$\beta_{NN}=0.5$					
3			$\beta_{VBD}=1.0$				
4				$\beta_{DT}=1.0$			
5					$\beta_{NN}=0.5$		
6						$\beta_{P}=1.0$	
7							$\beta_{NNS}=1.0$

- Fill diagonals with $\beta_j(k,k)$

Parse Triangle

	The (1)	gunman (2)	sprayed (3)	the (4)	building (5)	with (6)	bullets (7)
1	$\beta_{DT} = 1.0$	$\beta_{NP} = 0.25$					
2		$\beta_{NN} = 0.5$					
3			$\beta_{VBD} = 1.0$				
4				$\beta_{DT} = 1.0$			
5					$\beta_{NN} = 0.5$		
6						$\beta_{P} = 1.0$	
7							$\beta_{NNS} = 1.0$

- Calculate using induction formula

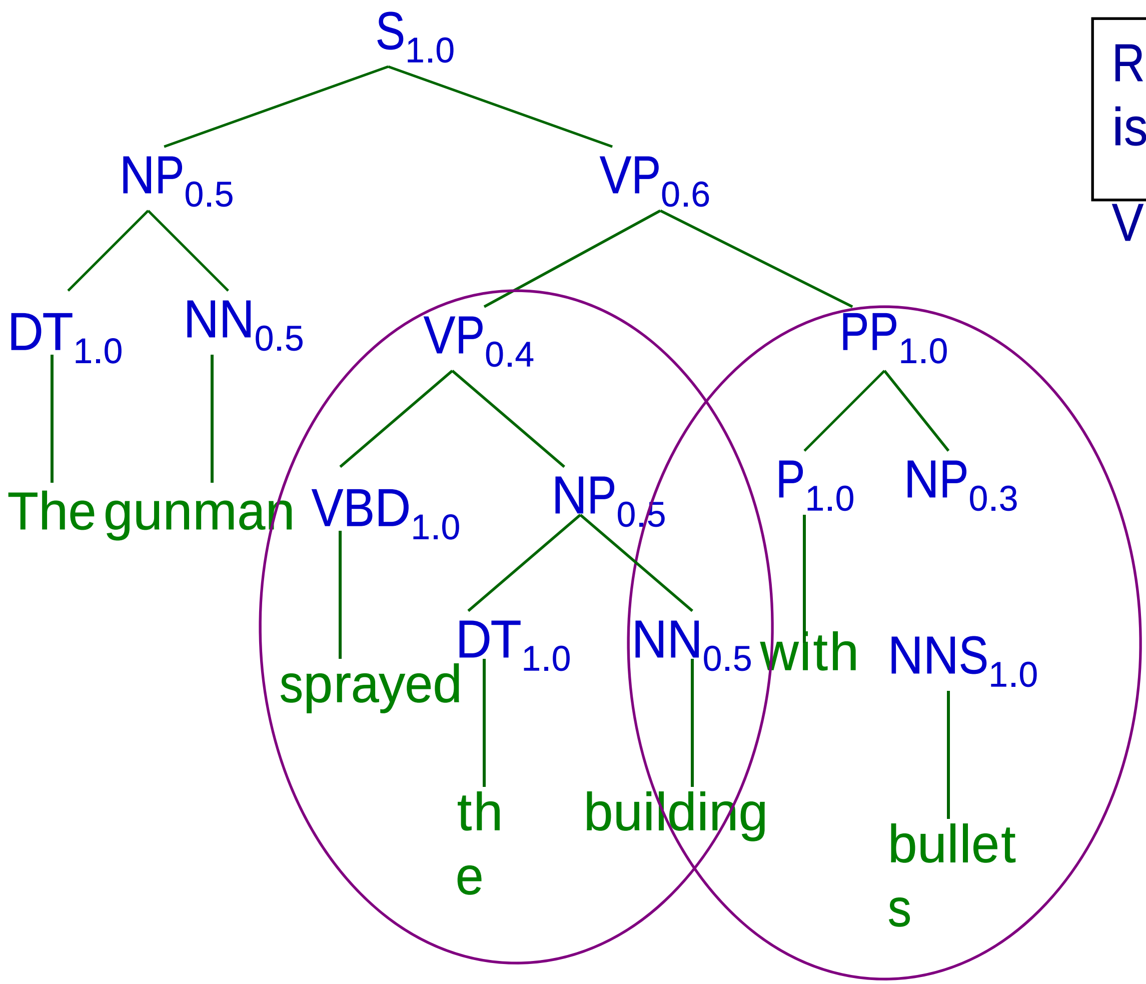
$$\beta_{NP}(1,2) = P(\text{the gunman} \mid NP_{1,2}, G)$$

$$= P(NP \rightarrow DT\ NN) * \beta_{DT}(1,1) * \beta_{NN}(2,2)$$

$$= 0.5 * 1.0 * 0.5 = 0.25$$

Example Parse t_1

- The gunman sprayed the building with bullets.

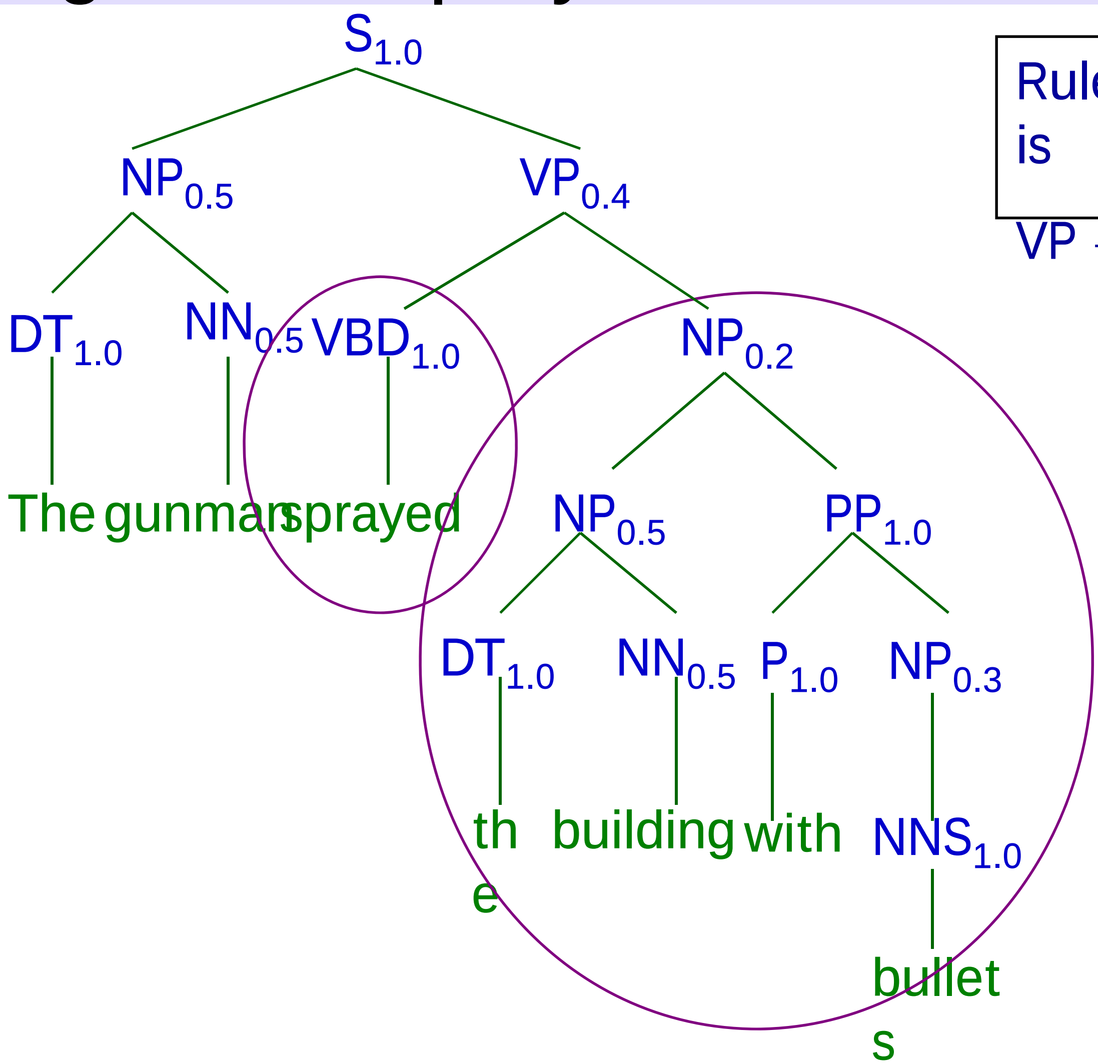


Rule used here is

VP → VP PP

Another Parse t_2

- The gunman sprayed the building with bullets.



Rule used here is

VP → VBD NP

Parse Triangle

	The (1)	gunman (2)	sprayed (3)	the (4)	building (5)	with (6)	bullets (7)
1	$\beta_{DT} = 1.0$	$\beta_{NP} = 0.25$					$\beta_{S} = 0.0465$
2		$\beta_{NN} = 0.5$					
3			$\beta_{VBD} = 1.0$		$\beta_{VP} = 1.0$		$\beta_{VP} = 0.186$
4				$\beta_{DT} = 1.0$	$\beta_{NP} = 0.25$		$\beta_{NP} = 0.015$
5					$\beta_{NN} = 0.5$		
6						$\beta_{P} = 1.0$	$\beta_{PP} = 0.3$
7							$\beta_{NNS} = 1.0$

$$\beta_{VP}(3,7) = P(\text{sprayed the building with bullets} \mid VP_{3,7}, G)$$

$$= P(VP \rightarrow VP\ PP) * \beta_{VP}(3,5) * \beta_{PP}(6,7)$$

$$+\ P(VP \rightarrow VBD\ NP) * \beta_{VBD}(3,3) * \beta_{NP}(4,7)$$

$$= 0.6 * 1.0 * 0.3 + 0.4 * 1.0 * 0.015 = 0.186$$

Different Parses

- Consider
  - Different splitting points :

    E.g., 5th and 3rd position
  - Using different rules for VP expansion :

    E.g., VP $\rightarrow$ VP PP, VP $\rightarrow$ VBD NP
- Different parses for the VP "sprayed the building with bullets" can be constructed this way.

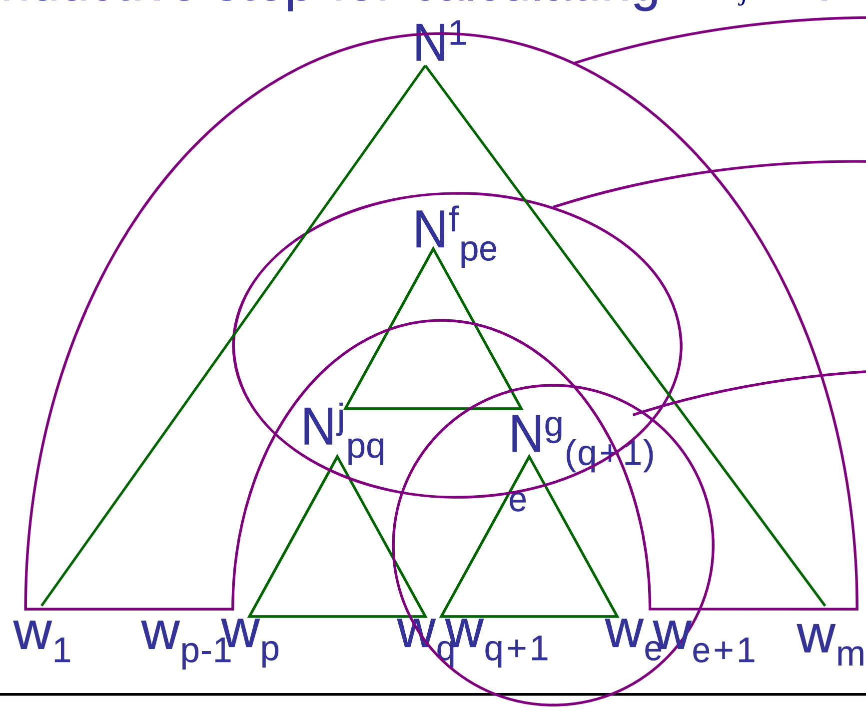
Outside Probabilities $\alpha_j(p,q)$

Base case:

$$\alpha_1(1,m) = 1 \text{ for start symbol}$$

$$\alpha_j(1,m) = 0 \text{ for } j \neq 1$$

Inductive step for calculating $\alpha_j(p,q)$:



$$\alpha_f(p,e)$$

$$P(N^f \rightarrow N^j N^g)$$

$$\beta_g(q+1,e)$$

Summation over f, g & e

Probability of a Sentence

$$P(w_{1m}, N_{pq} \mid G) = \sum_j P(w_{1m} \mid N^j_{pq}, G) = \sum_j \alpha_j(p,q)\beta_j(p,q)$$

- Joint probability of a sentence w_{1m} and that there is a constituent spanning words w_p to w_q is given as:

$$P(\text{The gunman....bullets}, N_{4,5} \mid G)$$

$$= \sum_j P(\text{The gunman...bullets} \mid N^j_{4,5}, G)$$

$$= \alpha_{NP}(4,5)\beta_{NP}(4,5)$$

$$+ \alpha_{VP}(4,5)\beta_{VP}(4,5) + \ldots$$