Natural Language Processing
CSE 325/425

Sihong Xie

Lecture 17:

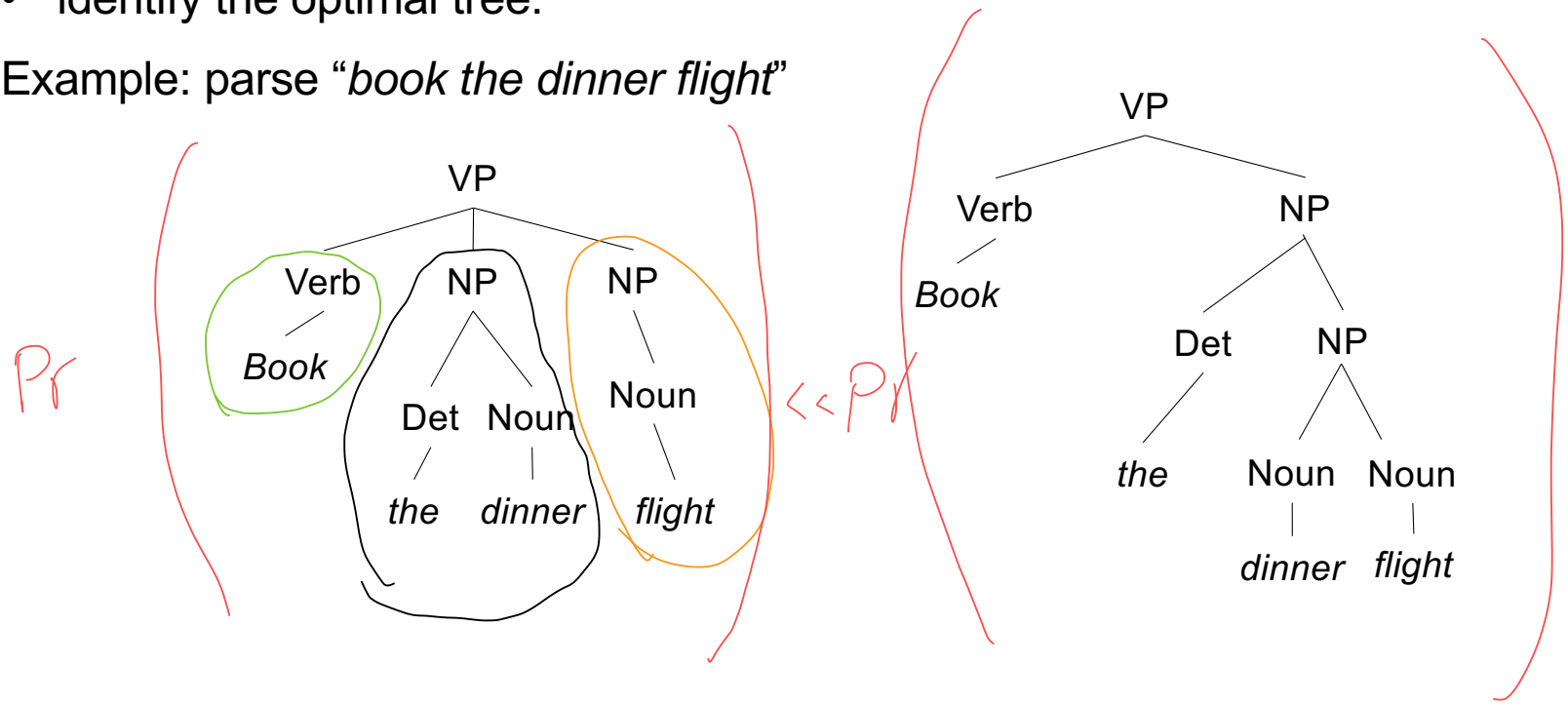
- Probabilistic CFG
- Inside probability and algorithm

Motivations

CYK algorithm only finds all parsing trees of a given sentence, and can't

- find probability of a tree;
- tell which tree is more likely than another trees;
- identify the optimal tree.

Example: parse "*book the dinner flight*"

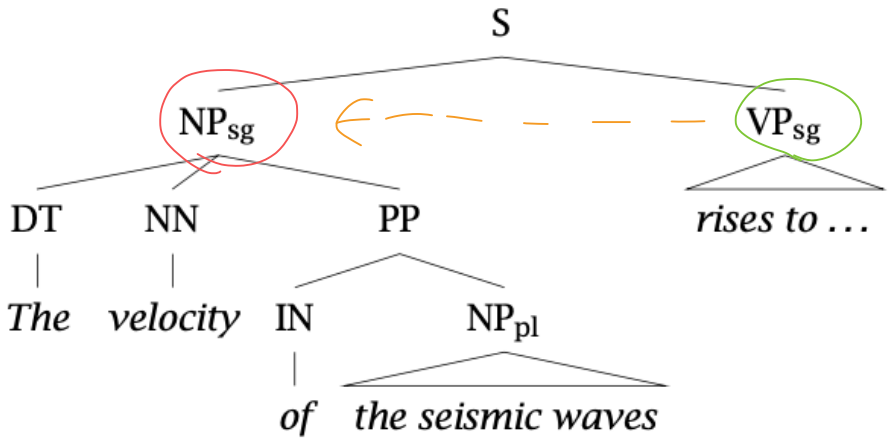


Motivations

HMM can only model linear relationship and has difficulty in long-range dependencies.

Example: parse "*The velocity of the seismic waves rises to …*"

- The singular form of the verb "rise" is attached to "*velocity*", not its predecessor.



CFG

PCFG

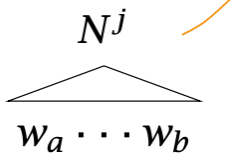
- Assign probability to each production in a CFG
  - the probabilities of the rules with the same left-hand-side sum to one.

Grammar G

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → *astronomers*	0.1
VP → V NP	0.7	NP → *ears*	0.18
VP → VP PP	0.3	NP → *saw*	0.04
P → *with*	1.0	NP → *stars*	0.18
V → *saw*	1.0	NP → *telescopes*	0.1

$\Pr(\text{NP PP} \mid \text{VP})$

  - Want to find the probability of any derivation $N^j \overset{*}{\Rightarrow} w_a \cdots w_b$



$$\Pr(N^j \overset{*}{\Rightarrow} w_a \ldots w_b | G)$$

  - In particular

$$\Pr(S \overset{*}{\Rightarrow} w_a \ldots w_b | G)$$

$$P\left(\begin{array}{c} {}^1\text{S} \\ {}^2\text{NP} \quad {}^3\text{VP} \\ \textit{the} \quad \textit{man} \quad \textit{snores} \end{array}\right)$$

A superscript labels the non-terminal
A subscript indicates the range of words covered

$$= P({}^1\text{S}_{13} \rightarrow {}^2\text{NP}_{12}\ {}^3\text{VP}_{33}, {}^2\text{NP}_{12} \rightarrow \textit{the}_1\ \textit{man}_2, {}^3\text{VP}_{33} \rightarrow \textit{snores}_3)$$

a b

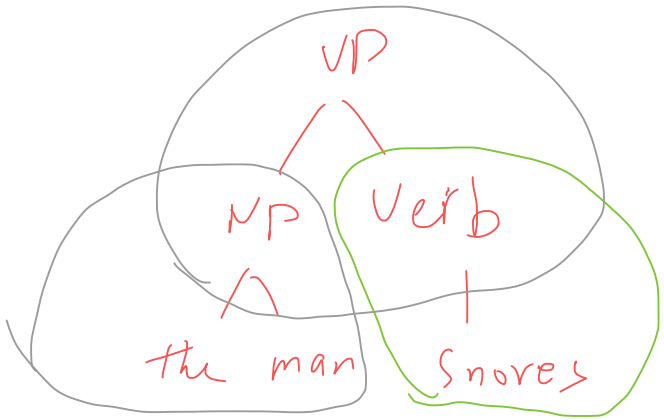
PCFG

- The probability of a parse tree *t* for the sentence $w_1 \ldots w_m$ given the grammar *G* is $\Pr(t, w_1 \ldots w_m | G)$ $= Pr(S \overset{*}{\Rightarrow} w_1 \cdots w_m \mid G)$
- The probability of the sentence $w_1 \ldots w_m$ is $\Pr(w_1 \ldots w_m | G) = \sum_t \Pr(t, w_1 \ldots w_m | G)$

Total Prob.

$P_r(x) = \sum_y Pr(x, y)$

- Properties:
  - Place invariance $\forall k \quad P(N^j_{k(k+c)} \rightarrow \zeta)$ is the same
  - Context-free $P(N^j_{kl} \rightarrow \zeta | \text{anything outside } k \text{ through } l) = P(N^j_{kl} \rightarrow \zeta)$
  - Ancestor-free $P(N^j_{kl} \rightarrow \zeta | \text{any ancestor nodes outside } N^j_{kl}) = P(N^j_{kl} \rightarrow \zeta)$



PCFG

- Probability of a parsing tree that deriving a sentence:

Example continued:

$$P\left(\begin{array}{ccc} & {}^1\mathrm{S} & \\ {}^2\mathrm{NP} & & {}^3\mathrm{VP} \\ \textit{the} & \textit{man} & \textit{snores} \end{array}\right)$$

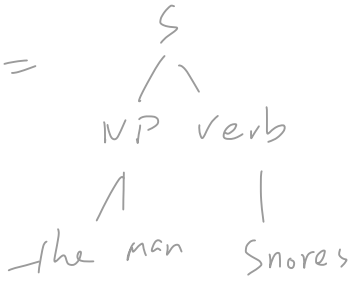
$$\begin{aligned}
&= P({}^1\mathrm{S}_{13} \rightarrow {}^2\mathrm{NP}_{12}\ {}^3\mathrm{VP}_{33}, {}^2\mathrm{NP}_{12} \rightarrow \textit{the}_1\ \textit{man}_2, {}^3\mathrm{VP}_{33} \rightarrow \textit{snores}_3) \\
&= P({}^1\mathrm{S}_{13} \rightarrow {}^2\mathrm{NP}_{12}\ {}^3\mathrm{VP}_{33}) P({}^2\mathrm{NP}_{12} \rightarrow \textit{the}_1\ \textit{man}_2 | {}^1\mathrm{S}_{13} \rightarrow {}^2\mathrm{NP}_{12}\ {}^3\mathrm{VP}_{33}) \\
&\quad P({}^3\mathrm{VP}_{33} \rightarrow \textit{snores}_3 | {}^1\mathrm{S}_{13} \rightarrow {}^2\mathrm{NP}_{12}\ {}^3\mathrm{VP}_{33}, {}^2\mathrm{NP}_{12} \rightarrow \textit{the}_1\ \textit{man}_2) \\
&= P({}^1\mathrm{S}_{13} \rightarrow {}^2\mathrm{NP}_{12}\ {}^3\mathrm{VP}_{33}) P({}^2\mathrm{NP}_{12} \rightarrow \textit{the}_1\ \textit{man}_2) P({}^3\mathrm{VP}_{33} \rightarrow \textit{snores}_3) \\
&= P(\mathrm{S} \rightarrow \mathrm{NP}\ \mathrm{VP}) P(\mathrm{NP} \rightarrow \textit{the man}) P(\mathrm{VP} \rightarrow \textit{snores})
\end{aligned}$$

def of conditional prob.

Ancestor free

place invariant.

$Pr(t, w_1 \cdots w_3) =$ where the tree



Three tasks with PCFG

- Find the probability of a sentence

$$\Pr(w_1 \dots w_m | G) = \sum_t \Pr(t, w_1 \dots w_m | G)$$

- Find the most likely parsing tree

$$t^* = \arg\max_t \Pr(t | w_1 \dots w_m, G)$$

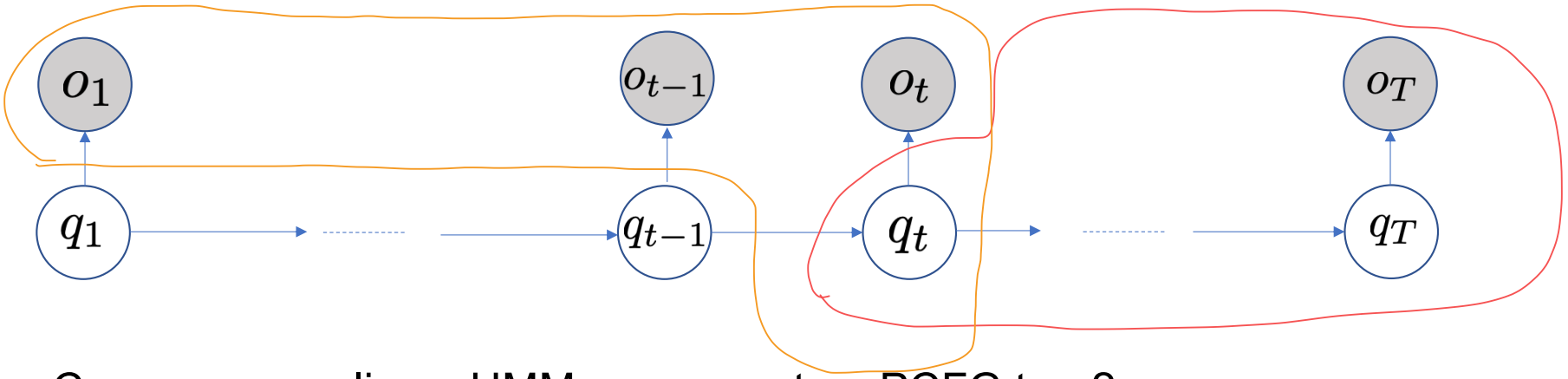
- Learn a PCFG from a training corpus using MLE.

$$G^* = \arg\max_G \Pr(w_1 \dots w_m | G)$$

From HMM to PCFG

- Assuming Chomsky Normal Form of the PCFG
- From HMM to PCFG

$$\alpha_t(j) = P(o_1, \ldots, o_t, q_t = j) \qquad \beta_t(j) = P(o_{t+1}, \ldots, o_T | q_t = j)$$



- Can we generalize a HMM sequence to a PCFG tree?

From HMM to PCFG

- Assuming Chomsky Normal Form of the PCFG
- From HMM to PCFG
- Between HMM and PCFG, there is a grammar called

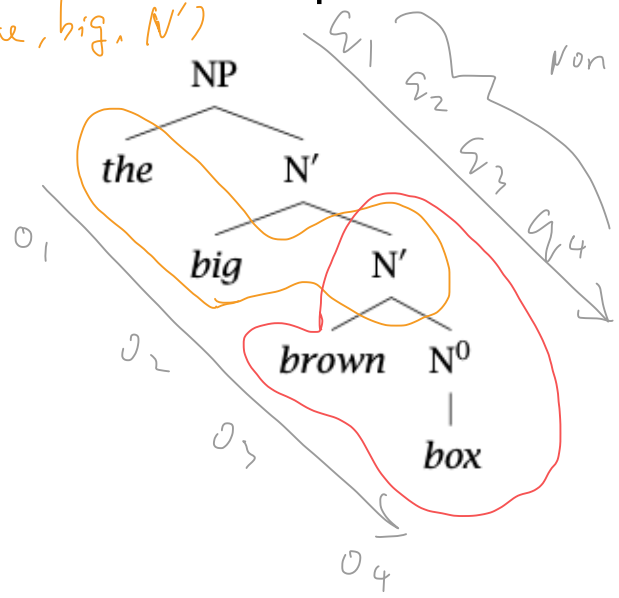
  Probabilistic Regular Grammar (PRG): $N^i \rightarrow w^j \; N^k$ or $N^i \rightarrow w^j$
- So what do the forward and backward probabilities correspond to on the tree?
  - Outside probability
  - Inside probability

forward

backward

Pr(the, big, N')

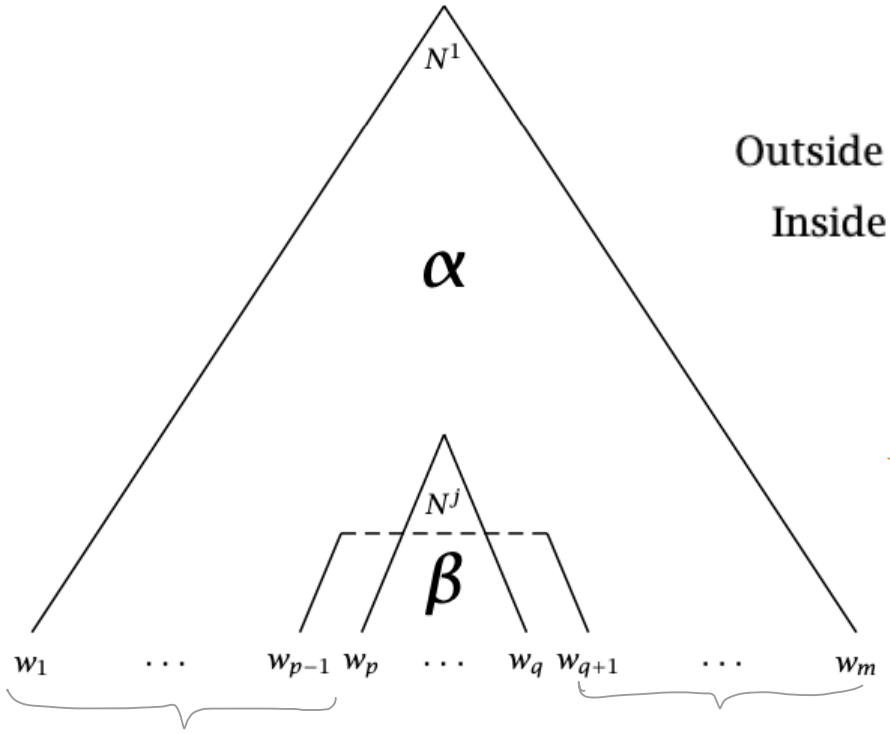
Pr(brown fox | N')



non-terminals

may not be POS-tags

Inside and outside probabilities



start end

$$\text{Outside probability} \quad \alpha_j(p,q) = P(w_{1(p-1)}, N^j_{pq}, w_{(q+1)m} | G)$$

$$\text{Inside probability} \quad \beta_j(p,q) = P(w_{pq} | N^j_{pq}, G)$$

$$\sum_j \alpha_j(p,q)\,\beta_j(p,q)$$

$$= \sum_j Pr(w_{1(p-1)}, N^j_{pq}, w_{(q+1)m} | G)$$

$$\times Pr(w_{pq} | N^j_{pq}, G)$$

$$= \sum_j Pr(w_{1m}, N^j_{pq} | G) = Pr(w_{1\cdots m} | G)$$

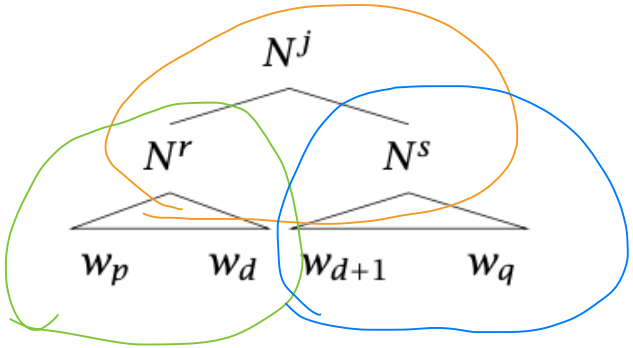
Inside probabilities

Sentence probability derived from a PCFG $P(w_{1m}|G) = P(N^1 \overset{*}{\Rightarrow} w_{1m}|G) = P(w_{1m}|N^1_{1m}, G) = \beta_1(1,m)$

- Base case

$$\begin{aligned}\beta_j(k,k) &= P(w_k|N^j_{kk}, G) \\ &= P(N^j \rightarrow w_k|G)\end{aligned}$$

- Induction due to CNF
  - CYK with probabilities



$$\begin{aligned}\beta_j(p,q) &= P(w_{pq}|N^j_{pq}, G) \\ &= \sum_{r,s}\sum_{d=p}^{q-1} P(w_{pd}, N^r_{pd}, w_{(d+1)q}, N^s_{(d+1)q}|N^j_{pq}, G) \\ &= \sum_{r,s}\sum_{d=p}^{q-1} P(N^r_{pd}, N^s_{(d+1)q}|N^j_{pq}, G) P(w_{pd}|N^j_{pq}, N^r_{pd}, N^s_{(d+1)q}, G) \\ &\quad \times P(w_{(d+1)q}|N^j_{pq}, N^r_{pd}, N^s_{(d+1)q}, w_{pd}, G) \\ &= \sum_{r,s}\sum_{d=p}^{q-1} P(N^r_{pd}, N^s_{(d+1)q}|N^j_{pq}, G) P(w_{pd}|N^r_{pd}, G) \\ &\quad \times P(w_{(d+1)q}|N^s_{(d+1)q}, G) \\ &= \sum_{r,s}\sum_{d=p}^{q-1} P(N^j \rightarrow N^r N^s)\beta_r(p,d)\beta_s(d+1,q)\end{aligned}$$

$\beta^{NP}(3,5) = 0.18 \times 0.18 \times 0.4$

Inside probabilities

Compute inside probabilities using dynamic programming

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → *astronomers*	0.1
VP → V NP	0.7	NP → *ears*	0.18
VP → VP PP	0.3	NP → *saw*	0.04
P → *with*	1.0	NP → *stars*	0.18
V → *saw*	1.0	NP → *telescopes*	0.1

$C(3,5) = \{ N^{j}_{35} \rightarrow N^{r}_{3d} \; N^{s}_{(d+1)5} \mid G \}$

$d=3 \quad N^{NP}_{35} \rightarrow N^{NP}_{33} \; N^{PP}_{45}$ $\quad (s = NP, r = NP, s = PP)$

$d=4 \quad N_{25} \rightarrow N_{24} \; N_{55}$

	1	2	3	4	5
1	$\beta_{NP} = 0.1$		$\beta_{S} = 0.0126$		$\beta_{S} = 0.0015876$
2		$\beta_{NP} = 0.04$ $\beta_{V} = 1.0$	$\beta_{VP} = 0.126$		$\beta_{VP} = 0.015876$
3			$\beta_{NP} = 0.18$		$\beta_{NP} = 0.01296$
4				$\beta_{P} = 1.0$	$\beta_{PP} = 0.18$
5					$\beta_{NP} = 0.18$
	astronomers	*saw*	*stars*	*with*	*ears*

← Textbook
FSNLP
Lehigh Lib
ebook

HW5: compute cell (1,3)