Lecture 7.5. HMM's to SCFG's

Note Title 1/26/2010

SCFG = $(\mathcal{V}, S, s, \mathcal{R})$ — Chomsky Normal Form

CFG:
- $\mathcal{V}$ ~ finite set of terminals V_0 {Sam, thinks, snores}
- S ~ finite set of non-terminals $S_0 = \{S, NP, VP, V\}$
- $\mathcal{R}$ - finite set of productions, $A \to BC \in S_0$, or $A \to X \in S_0$
- $s \in S$ is the start symbol ($s_0 = S$)

+ Production probabilities $p(A \to \beta) = P(\beta | A)$ for each $A \to \beta \in \mathcal{R}$

If $r_1 \dots r_n$ is sequence of productions used to generate a tree ψ, then

$$P_G(\psi) = p(r_1) \dots p(r_n) = \prod_{r \in \mathcal{R}} p(r)^{f_r(\psi)}$$

$$\sum_{\psi} P_G(\psi) = 1$$

$f_r(\psi)$ - no. times r is used.

What do we want to compute?

1. What is the prob. $P_G(\omega)$ of string ω

$$P_G(\omega) = P_G(S \Rightarrow^* \omega) = \sum_{\psi \in \Psi_G(\omega)} P_G(\psi)$$

2. What is the most probable parse $\hat{\psi}(\omega)$ of string ω?

$$\hat{\psi}(\omega) = \arg\max_{\psi \in \Psi_G(\omega)} P_G(\psi)$$

$\Psi_G(\omega)$ - set of parse trees for ω generated by G

$P_G(\psi)$ - prob of tree ψ w.r.t. grammar G.

SCFG "Inside" algorithm

generalization of DP

Goal: Compute $P(\omega) = \sum_{\psi \in \Psi_G(\omega)} P(\psi) = P(S \Rightarrow^* \omega)$

Data structure: table $P(A \Rightarrow^* \omega_{ij})$ for $A \in S$, $0 \le i < j \le n$

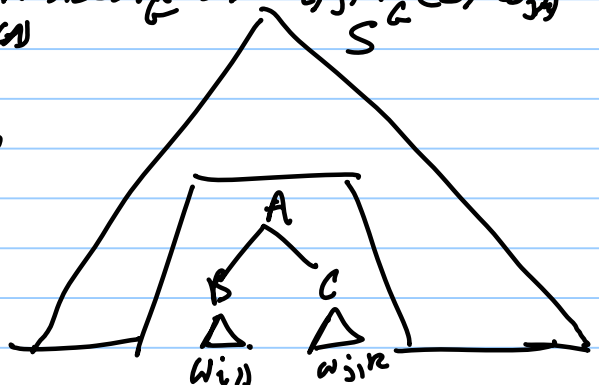
Base case: $P(A \Rightarrow^* \omega_{i,i+1}) = p(A \to \omega_i)$, $i = 0, \dots n-1$

Recursion: $P_G(A \Rightarrow^* \omega_{ik})$

$$= \sum_{j=i+1}^{k-1} \sum_{A \to BC \in \mathcal{R}(G)} p(A \to BC) P_G(B \Rightarrow^* \omega_{ij}) P_G(C \Rightarrow^* \omega_{jk})$$

Return: $P(S \Rightarrow^* \omega_{0,n})$

$P_G(A \Rightarrow^* \omega_{i,k})$ is called the "inside probability"



(2) Learning: Dataset of Examples

Assume the data is visible.

Count how frequently a rule is applied

Rule : $S \to NP\ VP$ Count 3 Frequency $\frac{3}{5}$

$NP \to Rice$ Count 2 " $\frac{2}{5}$

This is also the MLE for the SCFG
i.e. $P(\text{data} \mid \text{parameters } \theta)$

MLE is consistent asymptotically optimal (lowest variance)

But, this requires that the data is 'visible' or has been labelled. (like first HMM case)

Unsupervised Training EM (again)

$z = (x, y)$ x - observed data,
y - hidden variables (data)
θ - parameters

EM algorithm: given visible data x:

1. Guess initial value θ_0 of parameters (rule prob)
2. Repeat:

E-step: For all $y_1 \dots y_n \in Y$ generate pseudo-data $(x, y_1) \dots (x, y_n) \sim (x, y)$ 'frequency' $P_{\theta_i}(y \mid x)$

M-step: set θ_{i+1} to MLE from the pseudo-data

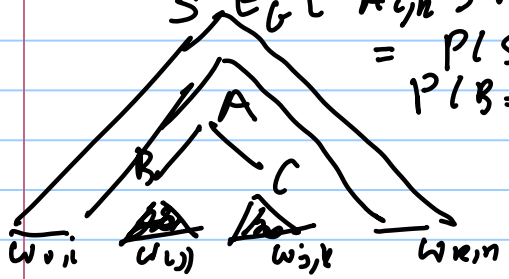
sometimes can perform maximization directly from sufficient statistics (expected production frequencies)

Use DP to calculate the expected probs:

$$E_G[n_{A \to BC} \mid w] = \sum_{0 \le i < j < k \le n} E_G[A_{i,k} \to B_{i,j} C_{j,k} \mid w]$$

expected fraction of parses of w in which $A_{i,k}$ rewrites as $B_{i,j}\ C_{j,k}$ is

$$E_G[A_{i,k} \to B_{i,j} C_{j,k} \mid w] = P(S \Rightarrow^* w_{1,i} A w_{k,n})\, p(A \to BC)\, P(B \Rightarrow^* w_{i,j})\, P(C \Rightarrow^* w_{j,k}) / P_A(w)$$



(3) Calculate $P_G(S \Rightarrow^* w_{0,i} A w_{k,n})$

"outside probabilities"

Recursion from larger to smaller substrings in w

Base case: $P(S \Rightarrow^* w_{0,0} S w_{n,n}) = ?$

Recursion: $P(S \Rightarrow^* w_{0,i} C w_{k,n}) =$

$$\sum_{j=1}^{i-1} \sum_{\substack{A,B \in S \\ A \to BC \in R}} P(S \Rightarrow^* w_{0,j} A w_{k,n}) p(A \to BC) P(B \Rightarrow^* w_{j,i})$$

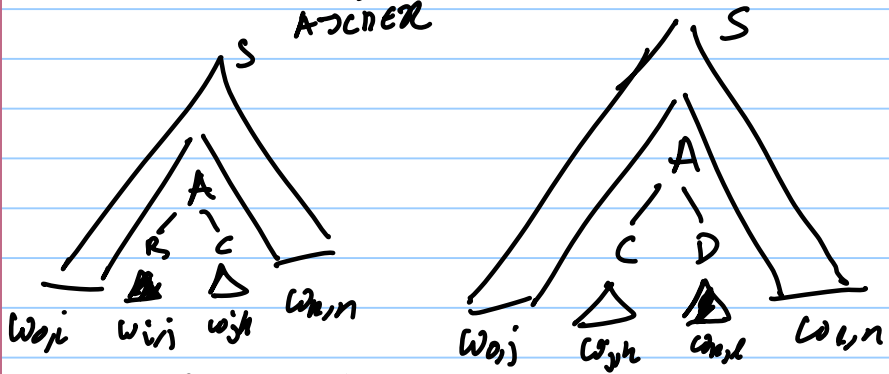
$$+ \sum_{l=k+1}^{n} \sum_{\substack{A,D \in S \\ A \to CD \in R}} P(S \Rightarrow^* w_{0,j} A w_{l,n}) p(A \to CD) P(D \Rightarrow^* w_{k,l})$$

Recursion in $P_G(S \Rightarrow^* w_{0,i} A w_{k,n})$

$$P(S \Rightarrow^* w_{0,j} C w_{k,n}) =$$

$$\sum_{i=0}^{k-1} \sum_{\substack{A,B \in S \\ A \to BC \in R}} P(S \Rightarrow^* w_{0,i} A w_{k,n}) p(A \to BC) P(B \Rightarrow^* w_{j,i})$$

$$+ \sum_{l=n+1}^{n} \sum_{\substack{A,D \in S \\ A \to CD \in R}} P(S \Rightarrow^* w_{0,j} A w_{l,n}) p(A \to CD) P(D \Rightarrow^* w_{n,l})$$



EM for SCFG:

Infer 'hidden structure' by ML of visible data.

1. guess initial rule probs.
2. Repeat:
   (a) parse a sample of sentences
   (b) weight each parse by its cond. prob
   (c) count rules used in each weighted parse, and estimate rule frequencies

- EM optimizes the likelihood of data D
- Each iteration is guaranteed not to decrease the likelihood of D. – but EM can get trapped in local minima
- The Inside-Outside algorithm can produce the expected counts without enumerating all parses of D

(4) Important to realize that EM follows from the Free Energy formulation

$-\log P(D|\theta)$ visible data

$D = \{x_1, \ldots x_m\}$

$P(D|\theta) = \prod_{i=m} P(x_i|\theta)$

$P(D|\theta) = P(D, y|\theta) = P(D|y,\theta)\, P(y)$

hidden

$\mathcal{F}[q,\theta] = \sum_y q(y) \log q(y) - \sum_y q(y) \log P(y, D|\theta)$

E-step: compute $q^{t+1}(y) = P(y|D,\theta^t)$

$= \frac{P(y, D|\theta^t)}{P(D|\theta^t)}$

M-step: compute

$\theta^{t+1} = \arg\min_\theta \{-\sum_y q^{t+1}(y) \log P(y, D|\theta)\}$.