Weighted Context-Free Grammars and Proper PCFGs

Mark Johnson

Brown University

4th August, 2005

based heavily on Zhiyi Chi (1999) "Statistical Properties of Context-Free Grammars", *Computational Linguistics*

Thanks to Noah Smith

Overview

- Not all PCFGs are *proper* (have a partition function $Z = 1$)
- Determining whether a PCFG is proper
- Maximum likelihood estimates of CFGs are always proper
- Weighted CFGs and Gibbs form CFGs
  - these arise naturally in conditional estimation
- WCFGs define the same distributions over trees as PCFGs
- how to convert a WCFG to an equivalent PCFG
- Conditional distributions
- CRF conditional distributions are HMM conditional distributions

Probabilistic Context-Free Grammars (1)

A PCFG is a tuple $G = (N, T, R, S, p)$ where:

- N is a finite set of nonterminals
- T is a finite set of terminals
- $R \subset N \times (N \cup T)^\star$ is a finite set of rules or productions
  - assume G does not contain useless rules or symbols
- $S \in N$ is the start symbol
- p is a function from $N \times (N \cup T)$ to $[0, 1]$ that is non-zero only on R. $p(A \to \alpha)$ is the probability of the rule $A \to \alpha$.

$$\text{for all } A \in N, \sum_{\alpha : A \to \alpha \in R} p(A \to \alpha) = 1$$

Probabilistic Context-Free Grammars (2)

The "probability" of a tree is the product of the probabilities of the rules used to generate it

$$p(t) = \prod_{A \to \alpha \in R} p(A \to \alpha)^{f_{A \to \alpha}(t)}$$

where:

- t is a parse tree and $p(t)$ is its "probability"
- $f_A(t)$ is the number of nodes labelled A in t
- $f_{A \to \alpha}(t)$ is the number of times $A \to \alpha$ is used in t

Not all PCFGs are proper

A PCFG is *proper* iff $\sum_{t \in \mathcal{T}} p(t) = 1$ where $\mathcal{T}$ is the set of all trees

$$R = \left\{ \begin{array}{ll} S \rightarrow S\ S & q \\ S \rightarrow a & 1-q \end{array} \right\}$$

$$\begin{aligned} Z_h &= \sum_{t:\text{height}(t) \leq h} p(t) \\ &= (1-q) + q s_{h-1}^2 \end{aligned}$$

so the fixed point $Z = \lim_{h \to \infty} Z_h = \sum_t p(t)$ satisfies

$$\begin{aligned} Z &= 1 - q + qZ^2 \\ Z &= \min(1, 1/q - 1) \\ &> 1 \text{ when } p > 1/2 \end{aligned}$$

Determining if a PCFG is proper

- Define a matrix M indexed by nonterminals in N

$$\begin{aligned} M_{A,B} &= \text{the expected number of } B\text{s that } A \text{ rewrites to} \\ &= \sum_{\alpha : A \to \alpha \in R} p(A \to \alpha) n_B(\alpha) \end{aligned}$$

  where $n_B(\alpha)$ is the number of Bs in α

- $M^k_{A,B}$ is the expected number of Bs that A rewrites to in k steps
- G is proper iff each $M^k_{A,B} \to 0$ as $k \to \infty$

  $\Leftrightarrow$ the largest eigenvalue of M is less than 1

- Wetherell (1980) describes an efficient way of determining this

MLEs of PCFGs are always proper (1)

- Relative frequency estimator from weighted set $\mathcal{D}$ of trees, where $W(t)$ is weight of tree t

$$p(A \to \alpha) \quad = \quad \frac{\sum_{t \in \mathcal{D}} f_{A \to \alpha}(t) W(t)}{\sum_{t \in \mathcal{D}} f_A(t) W(t)}$$

  – encompasses MLE from treebanks, and M step of EM

$$\begin{aligned}
q_A &= \mathrm{P}(A \text{ fails to terminate}) \\
&= \sum_{A \to \alpha} p(A \to \alpha) \mathrm{P}(\cup_i \{\alpha_i \text{ fails to terminate}\}) \\
&\leq \sum_{A \to \alpha} p(A \to \alpha) \sum_i \mathrm{P}(\{\alpha_i \text{ fails to terminate}\}) \\
&= \sum_{A \to \alpha} p(A \to \alpha) \sum_{B \in N} n_B(\alpha) q_B
\end{aligned}$$

MLEs of PCFGs are always proper (2)

$$\begin{aligned} q_A &\leq \sum_{A \to \alpha} p(A \to \alpha) \sum_B n_B(\alpha) q_B \\ &= \sum_B q_B \left(\frac{\sum_{A \to \alpha} n_B(\alpha) \sum_{t \in \mathcal{D}} f_{A \to \alpha}(t) W(t)}{\sum_{t \in \mathcal{D}} f_A(t) W(t)} \right) \end{aligned}$$

$$\begin{aligned} q_A \sum_{t \in \mathcal{D}} f_A(t) W(t) &\leq \sum_B q_B \sum_{t \in \mathcal{D}} \sum_{A \to \alpha} n_B(\alpha) f_{A \to \alpha}(t) W(t) \\ \sum_A q_A \sum_{t \in \mathcal{D}} f_A(t) W(t) &\leq \sum_B q_B \sum_{t \in \mathcal{D}} \sum_A \sum_{A \to \alpha} n_B(\alpha) f_{A \to \alpha}(t) W(t) \\ &= \sum_B q_B \sum_{t \in \mathcal{D}} \tilde{f}_B(t) W(t) \end{aligned}$$

where $\tilde{f}_B(t)$ is the number of *non-root* nodes labeled B in t.

Note that $f_S(t) = \tilde{f}_S(t) + 1$, and $f_A(t) = \tilde{f}_A(t)$ for all $A \neq S$

MLEs of PCFGs are always proper (3)

$$q_A \sum_{t \in \mathcal{D}} f_A(t) W(t) \;\; \leq \;\; \sum_B q_B \sum_{t \in \mathcal{D}} \tilde{f}_B(t) W(t)$$

$$\sum_A q_A \sum_{t \in \mathcal{D}} \left(f_A(t) - \tilde{f}_A(t) \right) W(t) \;\; \leq \;\; 0$$

But since $f_A(t) = \tilde{f}_A(t)$ for $A \neq S$ and $f_S(t) = \tilde{f}_S(t) + 1$:

$$q_S \sum_t W(t) \;\; \leq \;\; 0$$

which implies that $q_S = 0$

Weighted CFGs

- A *weighted CFG* G is one where each rule $A \to \alpha \in R$ is associated with a positive weight $w_{A\to\alpha}$. The weight $w(t)$ of a tree t generated by G is the product of the weights of the rules that generate it.

$$\begin{aligned} w(t) &= \prod_{A\to\alpha\in R} {w_{A\to\alpha}}^{f_{A\to\alpha}(t)} \\ Z &= \sum_{t\in\mathcal{T}} w(t) \\ \mathrm{P}(t) &= w(t)/Z \end{aligned}$$

- WCFGs can also be expressed as *Gibbs* or *log linear* models

$$\begin{aligned} \lambda_{A\to\alpha} &= \log w_{A\to\alpha} \text{ for each } A \to \alpha \in R \\ w(t) &= \prod_{A\to\alpha} \exp(\lambda_{A\to\alpha})^{f_{A\to\alpha}(t)} = \exp \sum_{A\to\alpha} \lambda_{A\to\alpha} f_{A\to\alpha}(t) \\ \mathrm{P}(t) &= \frac{1}{Z} \exp \sum_{A\to\alpha} \lambda_{A\to\alpha} f_{A\to\alpha}(t) \end{aligned}$$

Reasons for using WCFGs

- Unconstrained numerical optimization is easier than constrained numerical optimization
  $\Rightarrow$ numerically estimating a WCFG is easier than a PCFG
- Conditional estimation (trees given strings)
  - should be more accurate for parsing (given our bad grammars)
  - no closed form known (to me) $\Rightarrow$ numerical methods
  - there is a dynamic programming algorithm for calculating conditional likelihood and its derivatives
- Want to impose a prior on rule probabilities or regularize
  - Except for Dirichlet prior (which is conjugate to multinomial), numerical optimization probably required
  - Often easier to state priors/regularizers in log linear parameter space $\lambda_{A \to \alpha}$

Every WCFG dist is a proper PCFG dist

- In terms of grammars, WCFGs $\supset$ PCFGs $\supset$ proper PCFGs, but these all define exactly the same probability distributions over trees
- Chi 1999's rule probabilities for the equivalent proper PCFG:

$$
\begin{aligned}
p(A \to \alpha) &= \frac{1}{Z_A} w_{A\to\alpha} \prod_{k=1}^{|\alpha|} Z_{\alpha_k}, \text{ where} \\
Z_A &= \sum_{t \in \mathcal{T}_A} w(t) \\
\mathcal{T}_A &= \text{set of trees rooted in } A \\
\mathrm{P}_B(t) &= \frac{1}{Z_B} \prod_{A\to\alpha\in R} {w_{A\to\alpha}}^{f_{A\to\alpha}(t)} \text{ for } t \in \mathcal{T}_B
\end{aligned}
$$

$\Rightarrow$ many different WCFGs define the same distribution over trees

$\Rightarrow$ WCFGs weights are not identifiable even from treebank data

WCFG to PCFG conversion (1)

Proposition: $p_B(t) = \mathrm{P}_B(t)$ for all $B \in N$ and $t \in \mathcal{T}_B$, where

$$\begin{aligned} p(A \rightarrow \alpha) &= \frac{1}{Z_A} w_{A\rightarrow\alpha} \prod_{k=1}^{|\alpha|} Z_{\alpha_k} \\ p_B(t) &= \prod_{A\rightarrow\alpha} p(A \rightarrow \alpha)^{f_{A\rightarrow\alpha}(t)} \\ \mathrm{P}_B(t) &= \frac{1}{Z_B} \prod_{A\rightarrow\alpha} {w_{A\rightarrow\alpha}}^{f_{A\rightarrow\alpha}(t)} \end{aligned}$$

Proof by induction on the height h of t. Trivial for $h = 1$ (terminals)

WCFG to PCFG conversion (2)

Suppose $t \in \mathcal{T}_B$ has height $h > 1$, root rule is $B \to \beta$ and t_k is subtree rooted in kth child of t.

$$
\begin{aligned}
\mathrm{P}_B(t) &= \frac{1}{Z_B} \prod_{A \to \alpha} w_{A\to\alpha}{}^{f_{A\to\alpha}(t)} \\
&= \frac{1}{Z_B} w_{B\to\beta} \prod_{k=1}^{|\beta|} \prod_{A\to\alpha} w_{A\to\alpha}{}^{f_{A\to\alpha}(t_k)} \\
&= \frac{1}{Z_B} w_{B\to\beta} \prod_{k=1}^{|\beta|} Z_{\beta_k} \mathrm{P}_{\beta_k}(t_k) \\
&= \frac{1}{Z_B} w_{B\to\beta} \prod_{k=1}^{|\beta|} Z_{\beta_k} p_{\beta_k}(t_k) \text{ by induction hyp} \\
&= p(B \to \beta) \prod_{k=1}^{|\beta|} p_{\beta_k}(t_k) \\
&= p_B(t)
\end{aligned}
$$

Calculating WCFG partition functions Z_A

$$
\begin{aligned}
p(A \to \alpha) &= \frac{1}{Z_A} w_{A\to\alpha} \prod_{k=1}^{|\alpha|} Z_{\alpha_k} \\
\text{Let } \mathcal{T}_A^{(h)} &= \text{trees rooted in } A \text{ of height } \leq h \\
Z_A^{(h)} &= \sum_{t \in \mathcal{T}_A^{(h)}} \prod_{A\to\alpha} {w_{A\to\alpha}}^{f_{A\to\alpha}(t)} \\
Z_A &= \lim_{h\to\infty} Z_A^{(h)} \\
Z_A^{(1)} &= 1 \text{ if } A \in T,\ 0 \text{ otherwise} \\
Z_A^{(h+1)} &= \begin{cases} 1 & \text{for } A \in T \\ \displaystyle\sum_{\alpha: A\to\alpha\in R} w_{A\to\alpha} \prod_{k=1}^{|\alpha|} Z_{\alpha_k}^{(h)} & \text{for } A \in N \end{cases}
\end{aligned}
$$

Conditional Distributions

- Some WCFGs define *conditional* distributions of trees given strings, even though their partition functions Z diverge

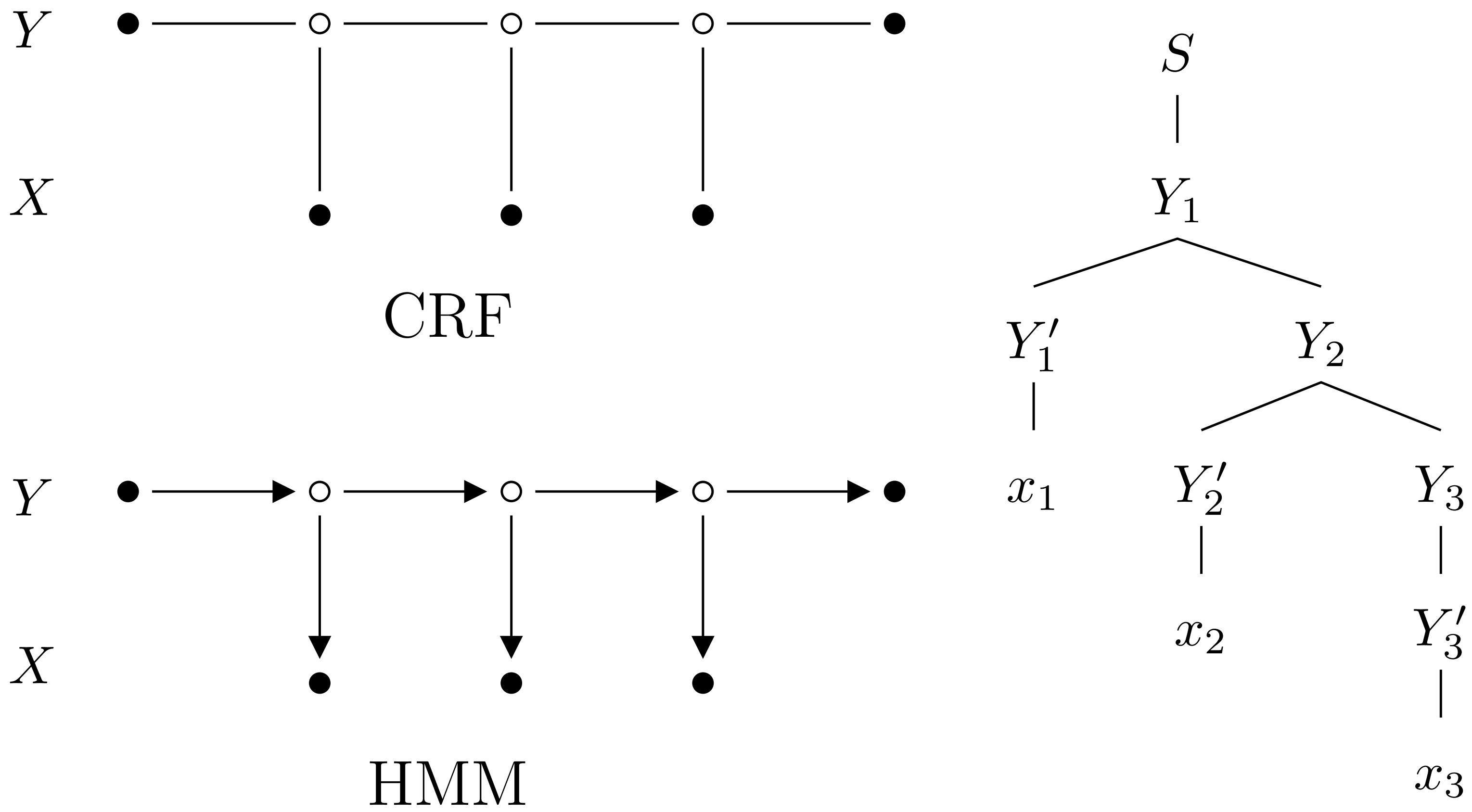
$$S \to S\,S; 1 \qquad S \to a; 1$$

- Chi's formula requires that all partition functions Z converge
- Change the WCFG weights $w_{A\to\alpha}$ to $\gamma^{|\alpha|-1} w_{A\to\alpha}$, $\gamma > 0$
  - multiplies the weight of a tree with yield y by $\gamma^{|y|-1}$

  $\Rightarrow$ doesn't affect the conditional distribution
- To find a PCFG with same conditional distribution as a WCFG
  1. Search for γ that makes the WCFG partition functions converge
  2. Apply the Chi formula to obtain PCFG rule probabilities

CRF cond dists = HMM cond dists

- A CRF is a WCFG $\Rightarrow \exists$ HMM with same cond dist



Conclusion

- Not all PCFGs are *proper* (have a partition function $Z = 1$)
- Determining whether a PCFG is proper
- Maximum likelihood estimates of CFGs are always proper
- Weighted CFGs and Gibbs form CFGs
  - these arise naturally in conditional estimation
- WCFGs define the same distributions over trees as PCFGs
- how to convert a WCFG to an equivalent PCFG
- Conditional distributions
- CRF conditional distributions are HMM conditional distributions