Probabilistic Context-Free Grammars

CFGs $\{N, T, S, R\}$

nonterminals, terminals, start, rules

S, NP, VP, PP

the, children, ate, cake, with, spoon

DT, NN, VBD, IN, NNS

POS: preterminals

S

binary

1 $S \rightarrow NP\ VP$

1/4 $VP \rightarrow VBD\ NP$

1/2 $NP \rightarrow DT\ NN$

1/2 $NP \rightarrow DT\ NNS$

unary

$VP \rightarrow VBD$ 3/4

$DT \rightarrow$ the

$NNS \rightarrow$ children

$NN \rightarrow$ cake

$NN \rightarrow$ spoon

$VBD \rightarrow$ ate

PCFG: rules have probs, probs sum to one per parent

$P(\text{rule} \mid \text{parent(rule)})$

$P(\text{tree } T) = \prod_{\text{rules}} P(r \mid \text{parent}(r))$

Steps to Parsing

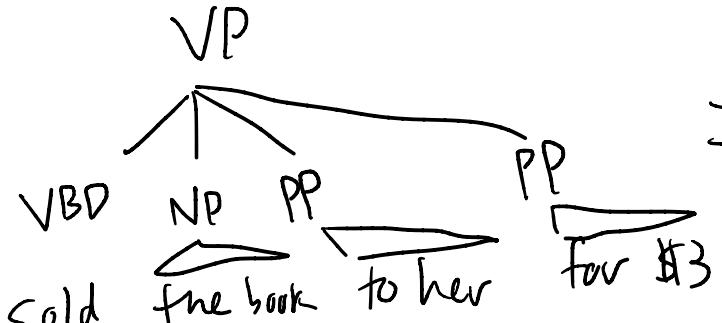
Treebank of sents labeled with trees

(1) Binarization

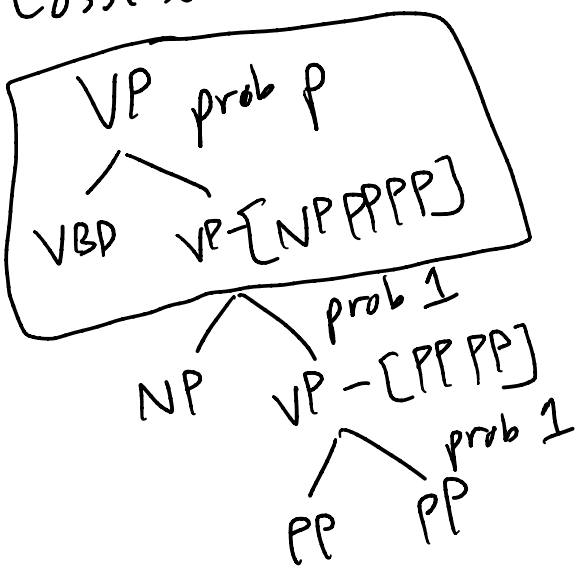
(2) Estimation of rule probs
(count + normalizing)
(maximum likelihood est.)

(3) Inference
$\text{argmax}_T P(T|\bar{x})$

Binarization



Lossless



Lossy