Natural Language Processing

CSE 325/425

Sihong Xie

Lecture 19:

- Training of PCFG
- Lexicalization of PCFG

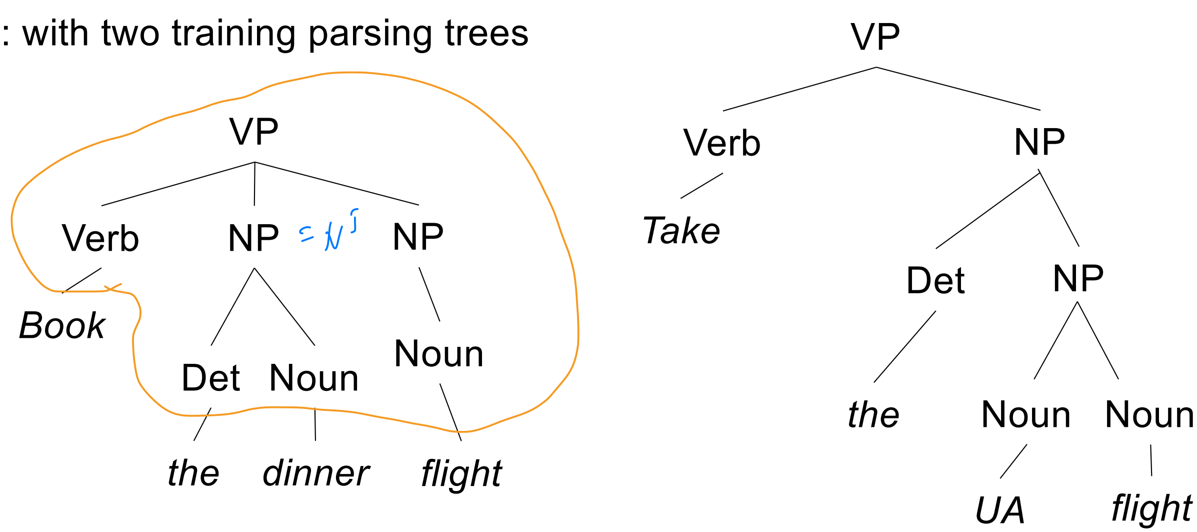
Training of PCFG

CYK with prob. on PCFG

f: Input $\rightarrow$ Output

- MLE: given a training corpus with sentences and their parsing trees (e.g., Penn Treebank),
  - extract rules observed in the corpus.
  - probability of a rule: how often it appears / how often the LHS appears.

Example: with two training parsing trees



$$Pr(NP \rightarrow Noun) = \frac{Count(NP \rightarrow Noun)}{Count(NP)} = \frac{1}{4}$$

Training of PCFG

What if we have a large number of sentences without being parsed?

- Need to deal with the unknown trees as latent structures;
- Similar to the unknown POS-tag sequences in the learning of HMM.
- EM algorithm. $\beta_j(p,q)$ $\alpha_j(p,q)$
  - E-step: run the inside and outside algorithms to find the inside and ~~output~~ outside probabilities. (Note: in HMM, this is the forward-backward algorithm).
  - M-step: estimate the rule probabilities based on the expectation of frequencies of occurrence using the inside/outside probabilities.

Training of PCFG

EM algorithm

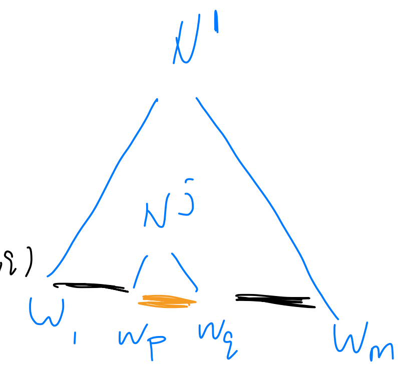
- Focus on the M-step, given the inside and outside probabilities.
  - Probability of a non-terminal N^j covering the range [p, q]:

$1 \le p \le q \le m$

$$P(N^j \overset{*}{\Longrightarrow} w_{pq} | N^1 \overset{*}{\Longrightarrow} w_{1m}, G) = \frac{\alpha_j(p,q)\beta_j(p,q)}{\pi}$$

← given

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



  - Expected frequency of the non-terminal:

$$E(N^j \text{ is used in the derivation}) = \sum_{p=1}^{m} \sum_{q=p}^{m} \frac{\alpha_j(p,q)\beta_j(p,q)}{\pi}$$

$\mathbb{E}_x[f(x)]$

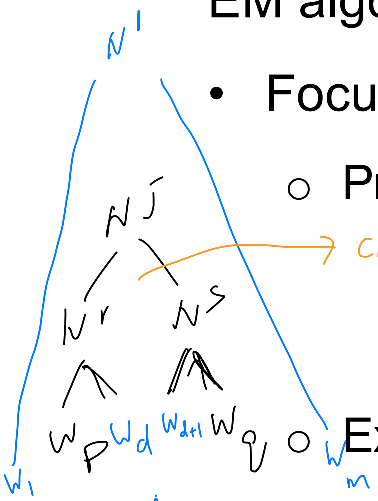
$= \sum_x P_r(x) f(x) = \sum_x P_r(x)$

$(f(x) = 1)$

$N^j \in N$ of PCFG extracted from some labeled corpus.

Training of PCFG

EM algorithm



CNF

d: unknown random var.

- Focus on the M-step, given the inside and outside probabilities.

dependent on previous est. of rule prob.

  - Probability of a rule deriving words in the range [*p, q*]:

$$P(N^j \to N^r\ N^s \stackrel{*}{\Longrightarrow} w_{pq} | N^1 \stackrel{*}{\Longrightarrow} w_{1m}, G)$$
$$= \boxed{\frac{\sum_{d=p}^{q-1} \alpha_j(p,q) P(N^j \to N^r\ N^s) \beta_r(p,d) \beta_s(d+1,q)}{\pi}}$$

  - Expected frequency of a rule:

$1 \le p \le q \le m$

$$E(N^j \to N^r\ N^s, N^j \text{ used})$$
$$= \frac{\sum_{p=1}^{m-1} \sum_{q=p+1}^{m} \boxed{\sum_{d=p}^{q-1} \alpha_j(p,q) P(N^j \to N^r\ N^s) \beta_r(p,d) \beta_s(d+1,q)}}{\pi}$$

$= \sum_p \sum_q P(N^j \to N^r N^s \stackrel{*}{\Rightarrow} w_{pq} | N^1 \stackrel{*}{\Rightarrow} w_{1m}, G)$

  - Estimation of the conditional probability of RHS given the LHS:

$$\hat{P}(N^j \to N^r\ N^s) = \frac{\sum_{p=1}^{m-1} \sum_{q=p+1}^{m} \sum_{d=p}^{q-1} \alpha_j(p,q) P(N^j \to N^r\ N^s) \beta_r(p,d) \beta_s(d+1,q)}{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_j(p,q) \beta_j(p,q)}$$

① from prev slide

② "π" is canceled out.

EM algorithm

EM algorithm

- Focus on the M-step, given the inside and outside probabilities.
  - Probability of the rule $N^j \rightarrow w^k$ deriving the word w^k somewhere in the sentence:

$$P(N^j \to w^k | N^1 \stackrel{*}{\Longrightarrow} w_{1m}, G) = \frac{\sum_{h=1}^{m} \alpha_j(h,h) P(N^j \to w_h, w_h = w^k)}{\pi}$$
$$= \frac{\sum_{h=1}^{m} \alpha_j(h,h) P(w_h = w^k) \beta_j(h,h)}{\pi}$$

  - Conditional probability of w^k given N^j:

$$\hat{P}(N^j \to w^k) = \frac{\sum_{h=1}^{m} \alpha_j(h,h) P(w_h = w^k) \beta_j(h,h)}{\sum_{p=1}^{m} \sum_{q=p}^{m} \alpha_j(p,q) \beta_j(p,q)}$$

from slide 4 and "π" is canceled out

Questioning the PCFG assumptions

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

- The probability of applying a rule is not independent of place

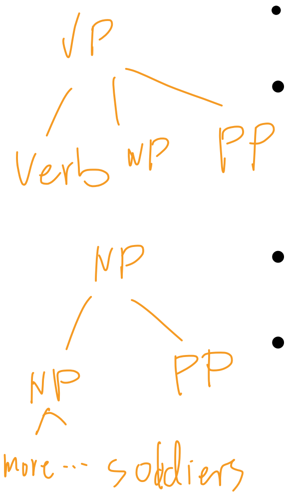
Subject
  - (a) **She's** able to take her baby to work with her. Pr(subject: NP->Pronoun) = 91%
  - (b) Uh, **my wife** worked until we had a family. Pr(subject: NP->Det Noun) = 9%

object
  - (a) Some laws absolutely prohibit **it**. Pr(object: NP->Pronoun) = 34%
  - (b) All the people signed **confessions**. Pr(object: NP->Det Noun) = 66%

Questioning the assumptions

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

- The probability of applying a rule is not independent of context (words)

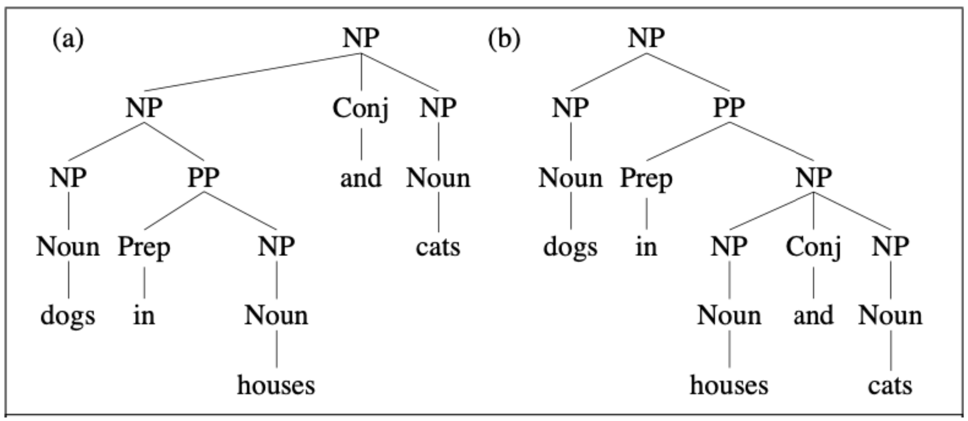


  - *"Moscow sent more than 100,000 soldiers into Afghanistan"*
  - should "*into Afghanistan*" be attached to "*sent*" (VP attachment) or "*more than 100,000 soldiers*" (NP attachment)?
  - Pr(NP attachment) > Pr(VP attachment) according to training corpora,
  - but in this sentence, it should be a VP attachment.

Questioning the assumptions

- Properties of PCFG:
  - 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 $P(N^j_{kl} \rightarrow \zeta | \text{any ancestor nodes outside } N^j_{kl}) = P(N^j_{kl} \rightarrow \zeta)$

- The probability of applying a rule is not independent of context (words)
  - Similar problem with conjunctions
  - "*dogs in houses and cats*"

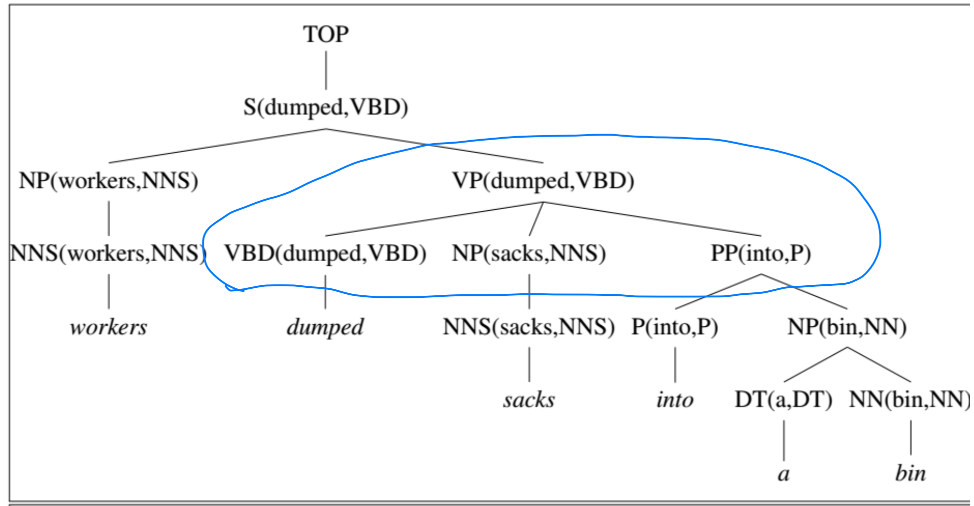


HW3: Pr(Tree 1)
= Pr(Tree 2) PCFG

VP → VBD NP PP
VP → VBD NP

Probabilistic lexicalized CFG

- Make the probability of a rule specific to the context (lexicons or words).
  - so that a parsing rule is selected depending on the contexts.
- A non-terminals in a rule adds:
  - head-word: the semantic center of the constituency;
  - head-POS-tag: how the head word is used.
- Copy the original ~~CFP~~ CFG multiple times for all possible lexicalizations.



Internal Rules				**Lexical Rules**		
TOP	→	S(dumped,VBD)		NNS(workers,NNS)	→	workers
S(dumped,VBD)	→	NP(workers,NNS)	VP(dumped,VBD)	VBD(dumped,VBD)	→	dumped
NP(workers,NNS)	→	NNS(workers,NNS)		NNS(sacks,NNS)	→	sacks
VP(dumped,VBD)	→	VBD(dumped, VBD)	NP(sacks,NNS) PP(into,P)	P(into,P)	→	into
PP(into,P)	→	P(into,P)	NP(bin,NN)	DT(a,DT)	→	a
NP(bin,NN)	→	DT(a,DT)	NN(bin,NN)	NN(bin,NN)	→	bin

Probabilistic lexicalized CFG

- Problem with estimating the probabilities of the lexicalized rules
  - *VP*(*dumped*,*VBD*)→*VBD*(*dumped*,*VBD*) *NP*(*sacks*,*NNS*) *PP*(*into*,*P*)

  - Pr(*VP*(*dumped*,*VBD*)→*VBD*(*dumped*,*VBD*) *NP*(*sacks*,*NNS*) *PP*(*into*,*P*))

    = Count of (*VP*(*dumped*,*VBD*)→*VBD*(*dumped*,*VBD*) *NP*(*sacks*,*NNS*) *PP*(*into*,*P*)) /

    Count of (*VP*(*dumped*,*VBD*))
  - How often, in the training corpus, can you see the expansion VP →VBD NP PP, and the head word is "*dumped*", which is POS-tagged as VBD, and the head word of NP is "*sacks*", which is POS-tagged as NNS, and the head word of PP is "*into*", which is POS-tagged as Preposition?
  - Data sparsity $\Longrightarrow$ *n*eed to make some independence assumption.
    - Same spirit as how HMM decomposes the joint probability of a long sequence.

$$Pr(\underset{q_1}{o} \to \underset{q_2}{o} \to \underset{q_3}{o} \to o \cdots \to o) = \prod_{t=1}^{T-1} Pr(q_{t+1} \mid q_t)$$

$$q_1 \perp\!\!\!\perp q_3 \mid q_2$$

STOP ← H → NP(Sacks, NNS) PP(into,P) STOP

Independent events

Collins Parser

- Specifies a particular way to calculate the problem of a rule
  - *VP(dumped,VBD)→VBD(dumped,VBD) NP(sacks,NNS) PP(into,P)*

    ⟹ *VP(dumped,VBD)→*STOP $\boxed{\textit{VBD(dumped,VBD)}}$ *NP(sacks,NNS) PP(into,P)* STOP
  - Pr(*VP(dumped,VBD)→* STOP *VBD(dumped,VBD) NP(sacks,NNS) PP(into,P)*) STOP

    = $\Pr_H$(*VBD|VP, dumped*) x

    $\Pr_L$(STOP |*VP, VBD, dumped*) x

    $\Pr_R$(*NP(sacks,NNS)* |*VP, VBD, dumped*) x

    $\Pr_R$(*PP(into,P)* |*VP, VBD, dumped*) x

    $\Pr_R$(STOP |*VP, VBD, dumped*)
  - Each of the above probability can be estimated using MLE
    - Less subject to data sparsity
    - since the joint events are more likely.