

# Tekniker för storskalig parsning

Generativa modeller

Joakim Nivre

Uppsala Universitet  
Institutionen för lingvistik och filologi  
`joakim.nivre@lingfil.uu.se`

# Generative Models

- ▶ A **generative** statistical model defines the **joint** probability  $P(x, y)$  of input  $x$  and output  $y$
- ▶ A **parameterization** defines  $P(x, y)$  as  $P(e_1, \dots, e_m)$ , where  $e_1, \dots, e_m$  (typically) correspond to substructures of  $x$  and  $y$
- ▶ Correct model is given by chain rule:

$$P(x, y) = P(e_1)P(e_2|e_1) \cdots P(e_m|e_1, \dots, e_{m-1})$$

- ▶ Tractable model requires **independence assumptions**

# PCFG as Generative Model

- ▶ Probability model of PCFG:

$$P(x, y) = \begin{cases} P(y) = \prod_{i=1}^{|R|} D(r_i)^{\text{count}(i,y)} & \text{if yield}(y) = x \\ 0 & \text{otherwise} \end{cases}$$

- ▶ Assumptions:

- ▶ Rule applications independent of each other:
  - ▶  $P(r_i = A \rightarrow \alpha_1 \cdots \alpha_k | r_1, \dots, r_{i-1}, r_{i+1}, \dots, r_m) = D(A \rightarrow \alpha_1 \cdots \alpha_k)$
- ▶ Joint probability of children given parent:
  - ▶  $P(B_1, \dots, B_m | A) = D(A \rightarrow B_1 \cdots B_m)$

## Example: NP Expansions in Penn Treebank

| Tree Context | NP PP | DT NN | PRP |
|--------------|-------|-------|-----|
| Anywhere     | 11%   | 9%    | 6%  |
| NP under S   | 9%    | 9%    | 21% |
| NP under VP  | 23%   | 7%    | 4%  |

- ▶ Pronominal realization (PRP) more frequent under S (subject)
- ▶ Prepositional modifiers more frequent under VP (object)

# History-Based Models

- ▶ History-based model:
  - ▶ Derivation of syntactic structure modeled by stochastic process
  - ▶ Process steps conditioned on events in the derivation history
- ▶ General form:

$$P(y) = \prod_{i=1}^m P(d_i | \Phi(d_1, \dots, d_{i-1}))$$

where

- ▶  $D = d_1, \dots, d_m$  is a derivation of  $y$
  - ▶  $\Phi$  is a function defining relevant events of the history
- ▶ PCFG as a history-based model:
  - ▶  $D =$  canonical derivation of  $y$  according to CFG
  - ▶  $\Phi(d_1, \dots, d_{i-1}) =$  left-hand side of rule used in  $d_i$

## Collins Model 2 [Collins 1997]

- ▶ Lexicalized nonterminals:  $A(a)$  ( $A \in N, a \in \Sigma$ )
- ▶ Expansion of  $A(a)$ :
  1. Choose a head child  $H$  with probability  $P_h(H|A, a)$
  2. Choose left and right subcat frames,  $LC$  and  $RC$ , with probabilities  $P_{lc}(LC|A, H, h)$  and  $P_{rc}(RC|A, H, h)$
  3. Generate the left and right modifiers  $L_1(l_1), \dots, L_k(l_k)$  and  $R_1(r_1), \dots, R_m(r_m)$  with  $P_l(L_i, l_i|A, H, h, \delta(i-1), LC)$  and  $P_r(R_i, r_i|A, H, h, \delta(i-1), RC)$
- ▶ Note:
  - ▶ Children generated inside-out from the head (with STOP)
  - ▶  $LC$  and  $RC$  are multisets of non-lexicalized nonterminals, deleted as the corresponding children are generated
  - ▶  $\delta(j)$  is a function of the surface string from the head word  $h$  to the outermost edge of the  $j$ th child on the same side:
    1. Is the string of zero length?
    2. Does the string contain a verb?
    3. Does the string contain 0, 1, 2 or more than 2 commas?

## Example: Collins Model 2

- ▶ VP in *Last week Marks bought Brooks*

$$\begin{aligned} P(S(\text{bought}) \rightarrow NP(\text{week}) NP\text{-C}(\text{Marks}) VP(\text{bought})) = & \\ & P_h(VP|S, \text{bought}) \times \\ & P_{lc}(\{NP\text{-C}\}|S, VP, \text{bought}) \times \\ & P_{rc}(\{\}\|S, VP, \text{bought}) \times \\ & P_l(NP\text{-C}(\text{Marks})|S, VP, \text{bought}, \langle 1, 0, 0 \rangle, \{NP\text{-C}\}) \times \\ & P_l(NP(\text{week})|S, VP, \text{bought}, \langle 0, 0, 0 \rangle, \{\}) \times \\ & P_l(\text{STOP}|S, VP, \text{bought}, \langle 0, 0, 0 \rangle, \{\}) \times \\ & P_r(\text{STOP}|S, VP, \text{bought}, \langle 0, 0, 0 \rangle, \{\}) \end{aligned}$$

# Properties of History-Based Models

- ▶ Lexicalization:
  - ▶ Nonterminals indexed by terminals (head child)
- ▶ Markovization:
  - ▶ Children generated one by one, conditioned on head/siblings
- ▶ History-based features:
  - ▶ Probabilities conditioned on top-down derivation
- ▶ Influential models:
  - ▶ Collins head-driven models [Collins 1997, Collins 1999]
  - ▶ Charniak's maximum-entropy inspired model [Charniak 2000]

# Parsing Model

- ▶ GEN( $x$ ):
  - ▶ Defined by (stochastic) system of derivations not necessarily constrained by a formal grammar
  - ▶ Number of candidate analyses in GEN( $x$ ) normally much larger than for treebank grammar
- ▶ EVAL( $y$ ):
  - ▶ Multiplicative model of joint probability  $P(x, y)$ , factored into  $P(d_i | \Phi(d_1, \dots, d_{i-1}))$  for each derivation step  $d_i$  given relevant parts of the derivation history

# Inference for History-Based Models

- ▶ In principle:
  - ▶ Use standard methods for PCFG parsing
- ▶ In practice:
  - ▶ Too inefficient (and memory intensive)
- ▶ Solutions:
  - ▶ Beam search: Only use  $k$  best items for each span [Collins 1997]
  - ▶ Coarse-to-fine parsing [Charniak 2000]:
    - ▶ First pass with simpler model (unlexicalized PCFG)
    - ▶ Second pass with full model, skipping low-probability items

# Learning for History-Based Models

- ▶ Maximum likelihood estimation:

$$\hat{P}(d_i | \Phi(d_1, \dots, d_{i-1})) = \frac{C[d_i, \Phi(d_1, \dots, d_{i-1})]}{C[\Phi(d_1, \dots, d_{i-1})]}$$

- ▶ Smoothing is critical
- ▶ Example:

$$\begin{aligned}\hat{P}(\text{NP}(n) | S, \text{VP}, v) &= \frac{C[\text{NP}(n), S, \text{VP}, v]}{C[S, \text{VP}, v]} \\ &= \frac{C[\text{NP}, S, \text{VP}, v]}{C[S, \text{VP}, v]} \\ &= \frac{C[\text{NP}, S, \text{VP}]}{C[S, \text{VP}]} \\ &= \frac{C[\text{NP}, \text{VP}]}{C[\text{VP}]}\end{aligned}$$

# Evaluation Criteria

- ▶ Robustness:
  - ▶ Yes, thanks to markovization (with proper smoothing)
- ▶ Disambiguation:
  - ▶ Yes, thanks to probability model
- ▶ Efficiency:
  - ▶ Reasonable with heavy pruning
- ▶ Accuracy:
  - ▶ Substantially higher than treebank PCFG [72% → 90%]

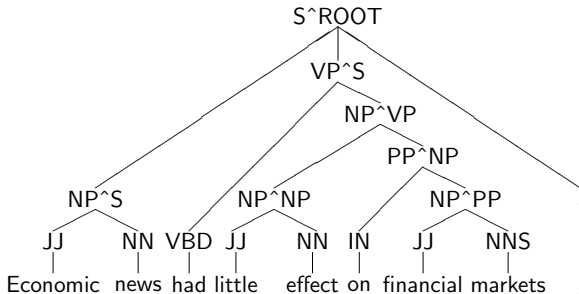
# PCFG Transformations

- ▶ From treebank PCFG to history-based model:
  - ▶ Retain syntactic representations, modify probability model
- ▶ Alternative:
  - ▶ Retain probability model, modify syntactic representations
  - ▶ Represent history-based model as PCFG
- ▶ Common transformations:
  - ▶ Parent annotation
  - ▶ State splitting
  - ▶ Binarization

# Parent Annotation

- ▶ Replace nonterminal  $A$  with  $A^B$  when  $A$  is child of  $B$

[Johnson 1998]



# State Splitting

- ▶ State splitting:
  - ▶ Split coarse linguistic categories into more fine-grained categories better suited for disambiguation
- ▶ Examples:
  - ▶ Lexicalization
  - ▶ Parent annotation
  - ▶ Subcategorization:
    - ▶ Manually constructed [Klein and Manning 2003]
    - ▶ Learned with latent variables and EM [Petrov et al. 2006]

# Binarization

- ▶ Replace  $n$ -ary grammar rule by a set of unary and binary rules
- ▶ Equivalent to markovization in history-based models
- ▶ Example:  $VP \rightarrow VB\ NP\ PP$ :

$$\begin{aligned} VP &\rightarrow \langle VP:VB \dots PP \rangle \\ \langle VP:VB \dots PP \rangle &\rightarrow \langle VP:VB \dots NP \rangle PP \\ \langle VP:VB \dots NP \rangle &\rightarrow \langle VP:VB \rangle NP \\ \langle VP:VB \rangle &\rightarrow VB \end{aligned}$$

# Parsing Model

- ▶ Treebank grammar over transformed trees/rules:
  - ▶  $G = (\Sigma, N, S, R, D)$
- ▶ GEN( $x$ ):
  - ▶ Defined by CFG  $G = (\Sigma, N, S, R)$
- ▶ EVAL( $x$ ) ( $y \in \text{GEN}(x)$ ):
  - ▶  $P(y)$  as defined by  $D$



# Inference

- ▶ Naive application of probabilistic CKY impossible
- ▶ Improvements:
  - ▶ Replace  $A(a) \in V$  by  $A \in V_{\text{base}} + \text{pointer to token } a$
  - ▶ Only consider (lexicalized) rules compatible with input words
- ▶ Effect:
  - ▶  $|R|$  is  $O(n^2)$  for sentence  $x = w_1, \dots, w_n$
  - ▶ Parsing complexity is  $O(n^5)$
- ▶ Still too inefficient to be practical:
  - ▶ Some kind of pruning of the search space is necessary
  - ▶ Same situation as for history-based models in general

# Learning

- ▶ Same learning methods as for treebank grammars:
  - ▶ Supervised learning: Relative frequency estimation
  - ▶ Unsupervised learning: Expectation-Maximization
- ▶ In addition:
  - ▶ State splits can be learned using latent variable models
  - ▶ Example:
    - ▶ Replace NP by NP(1), ..., NP( $k$ )
    - ▶ Learn which NP( $i$ ) to use in which rules using EM

# Evaluation Criteria

- ▶ Robustness:
  - ▶ Yes, with binarization and appropriate smoothing
- ▶ Disambiguation:
  - ▶ Yes, thanks to probability model
- ▶ Efficiency:
  - ▶ Reasonable with heavy pruning
- ▶ Accuracy:
  - ▶ Equivalent to history-based models

# Summary

- ▶ Generative statistical model:
  - ▶ Model of joint probability  $P(x, y)$  of input  $x$  and output  $y$
  - ▶ Parameterization  $P(x, y) = P(e_1) \cdots P(e_m)$
- ▶ Learning:
  - ▶ Relative frequency estimation from treebanks (MLE)
  - ▶ Expectation-maximization for hidden structure
- ▶ Inference:
  - ▶ Chart parsing with pruning for efficiency
  - ▶ Viterbi parsing (1-best) preserves complexity

## References and Further Reading

- ▶ Eugene Charniak. 2000.  
A maximum-entropy-inspired parser. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 132–139.
- ▶ Michael Collins. 1997.  
Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and the 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 16–23.
- ▶ Michael Collins. 1999.  
*Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.
- ▶ Mark Johnson. 1998.  
PCFG models of linguistic tree representations. *Computational Linguistics*, 24:613–632.
- ▶ Dan Klein and Christopher D. Manning. 2003.  
Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 423–430.
- ▶ Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006.

Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.