NLP
Introduction to NLP

Lexicalized Parsing 1
Limitations of PCFGs

• The probabilities don’t depend on the specific words
  – E.g., *give* someone something (2 arguments) vs. *see* something (1 argument)

• It is not possible to disambiguate sentences based on semantic information
  – E.g., eat pizza with *pepperoni* vs. eat pizza with *fork*

• Lexicalized grammars – idea
  – Use the head of a phrase as an additional source of information
  – VP[ate] → V[ate]
  – Fundamental idea in syntax, cf. X–bar theory, HPSG
  – Constituents receive their heads from their head child
Head Extraction Example (Collins)

- NP -> DT NNP NN  (rightmost)
- NP -> DT NN NNP  (rightmost)
- NP -> NP PP       (leftmost)
- NP -> DT JJ       (rightmost)
- NP -> DT          (rightmost leftover child)
Collins Parser (1997) 1/2

• Generative, lexicalized model
• Horizontal markovization
  – only condition on the head (also on the distance $\Delta$ from the head)
• Types of rules
  – LHS $\rightarrow$ $L_n L_{n-1} \ldots L_1 H \ R_1 \ldots R_{m-1} R_m$
  – H gets generated first
  – L gets generated next
  – R gets generated last
• Maximum likelihood estimates

$$P_{ML} (\text{PPof-IN} \mid \text{VPthink-VB}) =$$
$$\frac{\text{Count (PPof-IN right of the head VPthink-VB)}}{\text{Count (symbols right of the head VPthink-VB)}}$$

• Smoothing (lexicalized, unlexicalized, “unheaded”)

$$\text{smoothedP (PPof-IN} \mid \text{VPthink-VB}) = \lambda_1 P (\text{PPof-IN} \mid \text{VPthink-VB}) +$$
$$+ \lambda_2 P (\text{PPof-IN} \mid \text{VP-VB}) + (1-\lambda_1-\lambda_2) P (\text{PPof-IN} \mid \text{VP})$$
Issues with Lexicalized Grammars

• Sparseness of training data
  – Many probabilities are difficult to estimate from the Penn Treebank
  – E.g., WHADJP (when not “how much” or “how many” only appears 6 times out of 1M constituents)
  – Smoothing is essential

• Combinatorial explosion
  – Parameterization is essential
Discriminative Reranking

• Issues with statistical parsers
  – A parser may return many parses of a sentence, with small differences in probabilities
  – The top returned parse may not necessarily be the best because the PCFG may be deficient
• Other considerations may need to be taken into account
  – parse tree depth
  – left attachment vs. right attachment
  – discourse structure
• Can you think of others features that may affect the reranking?
Answer

• Considerations that may affect the reranking
  – parse tree depth
  – left attachment vs. right attachment
  – discourse structure

• Can you think of others?
  – consistency across sentences
  – or other stages of the NLU pipeline
Discriminative Reranking

• n–best list
  – Get the parser to produce a list of n–best parses (where n can be in the thousands)

• Reranking
  – Train a discriminative classifier to rerank these parses based on external information such as a bigram probability score or the amount of right branching in the tree
Statistical Parser Performance

- F1 (sentences <= 40 words)
  - Charniak (2000) – 90.1%
  - Charniak and Johnson (2005) – 92% (discriminative reranking)
Notes

• Complexity of lexicalized parsing
  – $O(N^5 g^3 V^3)$, instead of $O(N^3)$ because of the lexicalization
    • $N$ = sentence length
    • $g$ = number of non-terminals
    • $V$ = vocabulary size
  – Use beam search (Charniak; Collins)

• Sparse data
  – 40,000 sentences; 12,409 rules (Collins)
  – 15% of all test sentences contain a rule not seen in training (Collins)
Notes

• Complements (arguments) vs. adjuncts (additional information)
  – NP–C (Collins)

• Subcategorization
  – E.g., transitive vs. intransitive verbs

• Parent annotation
  – NP^S (Johnson 1998)
Another Example

$P_h(V \mid VP, \text{told}, V) \times$

$P_{lc}({\{} \mid VP, V, \text{told}, V) \times$

$P_d(\text{STOP} \mid VP,V,\text{told},V,\text{LEFT},\Delta = 1,{\{})) \times$

$P_{rc}([\text{NP-C, SBAR-C}] \mid VP, V, \text{told, V}) \times$

$P_d(\text{NP-C}(\text{Bill,NNP}) \mid VP,V,\text{told},V,\text{RIGHT},\Delta = 1,\{\text{NP-C, SBAR-C}\}) \times$

$P_d(\text{NP}(\text{yesterday,NN}) \mid VP,V,\text{told},V,\text{RIGHT},\Delta = 0,\{\text{SBAR-C}\}) \times$

$P_d(\text{SBAR-C}(\text{that,COMP}) \mid VP,V,\text{told},V,\text{RIGHT},\Delta = 0,\{\text{SBAR-C}\}) \times$

$P_d(\text{STOP} \mid VP,V,\text{told},V,\text{RIGHT},\Delta = 0,\{\})$
Notes

• Learning PCFG without an annotated corpus
  – Use EM (inside–outside) (Baker 1979), limited success

• Summary
  – Lexicalization takes F1 from 75% to 90%

• Markovization
  – Horizontal (forgetful binarization)
  – Vertical (generalized parent annotation)
    • Note: infinite vertical markovization is inefficient (Klein and Manning 2003)

• Collins and Charniak are generative models
• Reranking is a discriminative model
NLP