Seven Lectures on Statistical Parsing

Christopher Manning
LSA Linguistic Institute 2007
LSA 354
Lecture 5

A bit more on lexicalized parsing

Complexity of lexicalized PCFG parsing

Time charged:
- \( i, k, j \Rightarrow n^3 \)
- \( A, B, C \Rightarrow g^3 \)
- Naively, \( g \) becomes huge
- \( d_1, d_2 \Rightarrow n^2 \)

Running time is \( O(g^3 \times n^5) \)!!

Complexity of exhaustive lexicalized PCFG parsing

Refining the node expansion probabilities

- Charniak (1997) expands each phrase structure tree in a single step.
- This is good for capturing dependencies between child nodes
- But it is bad because of data sparseness
- A pure dependency, one child at a time, model is worse
- But one can do better by in between models, such as generating the children as a Markov process on both sides of the head (Collins 1997; Charniak 2000)
- Cf. the accurate unlexicalized parsing discussion
Collins (1997, 1999); Bikel (2004)

- Collins (1999): also a generative model
- Underlying lexicalized PCFG has rules of form
  \[ P \rightarrow L_1 L_{i-1} \cdots L_i H R_i \cdots R_{i+1} R_1 \]
- A more elaborate set of grammar transforms and factorizations to deal with data sparseness and interesting linguistic properties
- So, generate each child in turn: given \( P \) has been generated, generate \( H \), then generate modifying nonterminals from head-adjacent outward with some limited conditioning

Overview of Collins’ Model

Collins model … and linguistics

- Collins had 3 generative models: Models 1 to 3
- Especially as you work up from Model 1 to 3, significant linguistic modeling is present:
  - Distance measure: favors close attachments
  - Model is sensitive to punctuation
  - Distinguish base NP from full NP with post-modifiers
  - Coordination feature
  - Mark gapped subjects
  - Model of subcategorization; arguments vs. adjuncts
  - Slash feature/gap threading treatment of displaced constituents
  - Didn’t really get clear gains from this.

Collins (1997): also a generative model

• Underlying lexicalized PCFG has rules of form

\[ P \rightarrow L_1 L_{i-1} \cdots L_i H R_i \cdots R_{i+1} R_1 \]

• A more elaborate set of grammar transforms and factorizations to deal with data sparseness and interesting linguistic properties

• So, generate each child in turn: given \( P \) has been generated, generate \( H \), then generate modifying nonterminals from head-adjacent outward with some limited conditioning

Modifying nonterminals generated in two steps

\[ P_M \quad S(VBD\rightarrow \text{sat}) \]
\[ NP(NNP\rightarrow \text{John}) \]
\[ VP(VBD\rightarrow \text{sat}) \]

Smoothing for head words of modifying nonterminals

<table>
<thead>
<tr>
<th>Back-off level</th>
<th>( P_w(\text{subcat}_w) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( M_w, A, \text{coord, punct, } P, H, ) ( A_w, \text{subcat}_w )</td>
</tr>
<tr>
<td>1</td>
<td>( M_w, A, \text{coord, punct, } P, H, A, A_w, \text{subcat}_w )</td>
</tr>
<tr>
<td>2</td>
<td>( I_w )</td>
</tr>
</tbody>
</table>

- Other parameter classes have similar or more elaborate backoff schemes

Collins model … and linguistics

- Collins had 3 generative models: Models 1 to 3
- Especially as you work up from Model 1 to 3, significant linguistic modeling is present:
  - Distance measure: favors close attachments
  - Model is sensitive to punctuation
  - Distinguish base NP from full NP with post-modifiers
  - Coordination feature
  - Mark gapped subjects
  - Model of subcategorization; arguments vs. adjuncts
  - Slash feature/gap threading treatment of displaced constituents
  - Didn’t really get clear gains from this.

Bilexical statistics: Is use of maximal context of \( P_M \) useful?

- Collins (1999): “Most importantly, the model has parameters corresponding to dependencies between pairs of headwords.”
- Gildea (2001) reproduced Collins’ Model 1 (like regular model, but no subcats)
  - Removing maximal back-off level from \( P_M \) resulted in only 0.5% reduction in F-measure
  - Gildea’s experiment somewhat unconvincing to the extent that his model’s performance was lower than Collins’ reported results
Choice of heads

- If not bilexical statistics, then surely choice of heads is important to parser performance...
- Chiang and Bikel (2002): parsers performed decently even when all head rules were of form "if parent is X, choose left/rightmost child"
- Parsing engine in Collins Model 2-emulation mode: LR 88.55% and LP 88.80% on §Ø0 (sent. len. ≤ 40 words)
  - compared to LR 89.9%, LP 90.1%

Use of maximal context of $P_{MW}$

  - There was nothing maximum entropy about it. It was a cleverly smoothed generative model
  - Smooths estimates by smoothing ratio of conditional terms (which are a bit like maxent features):
    \[
    \frac{P(t|I_p,t_p)}{P(t|I_p)}
    \]
  - Biggest improvement is actually that generative model predicts head tag first and then does $P(w_t,\ldots)$
    - Like Collins (1999)
    - Markovizes rules similarly to Collins (1999)
    - Gets 90.1% LP/LR F score on sentences ≤ 40 wds

  - Can you automatically find good symbols?
    - Brackets are known
    - Base categories are known
    - Induce subcategories
    - Clever split/merge category refinement
  - EM algorithm, like Forward-Backward for HMMs, but constrained by tree.

Use of maximal context of $P_{MW}$

- Full model
  - LR 89.9, LP 90.1
  - 0 CBs: 68.8, 7 CBs: 89.2, 1 CBs: 88.8

Bilexical statistics are used often

- The 1.4% use of bilexical dependencies suggests they don’t play much of a role in parsing
  - But the parser pursues many (very) incorrect theories
- So, instead of asking how often the decoder can use bigram probability on average, ask how often while pursuing its top-scoring theory
- Answering question by having parser constrain parse its own output
  - train as normal on §§02–21
  - parse §Ø0
  - feed parse trees as constraints
- Percentage of time parser made use of bigram statistics shot up to 28.8%
- So, used often, but use barely affect overall parsing accuracy
  - Exploratory Data Analysis suggests explanation
    - distributions that include head words are usually sufficiently similar to those that do not as to make almost no difference in terms of accuracy

Performance on §Ø0 of Penn Treebank on sentences of length ≤ 40 words
Number of phrasal subcategories

POS tag splits, commonest words: effectively a class-based model

- Proper Nouns (NNP):
  - NNP-15: Wall
  - NNP-14: San
  - NNP-13: New
  - NNP-2: Noriega
  - NNP-1: Bush
  - NNP-0: L.

- Personal Pronouns (PRP):
  - PRP-0: it
  - PRP-1: he
  - PRP-2: them

The Latest Parsing Results...

<table>
<thead>
<tr>
<th>Parser</th>
<th>F1 ≤ 40 words</th>
<th>F1 all words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klein &amp; Manning unlexicalized 2003</td>
<td>86.3</td>
<td>85.7</td>
</tr>
<tr>
<td>Matsuzaki et al. simple EM latent states 2005</td>
<td>86.7</td>
<td>86.1</td>
</tr>
<tr>
<td>Charniak generative (“maxent inspired”) 2000</td>
<td>90.1</td>
<td>89.5</td>
</tr>
<tr>
<td>Petrov and Klein NAAACL 2007</td>
<td>90.6</td>
<td>90.1</td>
</tr>
<tr>
<td>Charniak &amp; Johnson discriminative reranker 2005</td>
<td>92.0</td>
<td>91.4</td>
</tr>
</tbody>
</table>

Treebanks and linguistic theory

Treebanks

- Treebank and parsing experimentation has been dominated by the Penn Treebank
- If you parse other languages, parser performance generally heads, downhill, even if you normalize for treebank size:
  - WSJ small: 82.5%
  - Chinese TB: 75.2%
- What do we make of this? We’re changing several variables at once.
  - “Is it harder to parse Chinese, or the Chinese Treebank?” [Levy and Manning 2003]
- What is the basis of the structure of the Penn English Treebank anyway?
Treebanks

- A good treebank has an extensive manual for each stage of annotation
  - Consistency is often at least as important as being right
  - But is what’s in these treebanks right?

- Tokenization:
  - has n’t / ‘n’t
- Hyphenated terms
  - cancer-causing JJ asbestos/NN
  - the back-on-terra-firma JJ toast/NN
  - the DT nerd-and-geek JJ club/NN

Treebank: POS

- Some POS tagging errors reflect not only human inconsistency but problems in the definition of POS tags, suggestive of clines/blends
  - Example: near
  - In Middle English, an adjective
  - Today is it an adjective or a preposition?
    - The near side of the moon
    - We were near the station
  - Not just a word with multiple parts of speech!
  - Evidence of blending:
    - I was nearer the bus stop than the train

- In some cases functional/notional tagging dominates structure in Penn Treebank, even against explicit instructions to the contrary:
  - worth: 114 instances
    - 10 tagged IN (8 placed in ADJP)
    - 65 tagged JJ (48 in ADJP, 13 in PP, 4 NN/NP errors)
    - 39 tagged NN (2 IN/JJ errors)
  - Linguist hat on: I tend to agree with IN choice (when not a noun):
    - tagging accuracy only 41% for worth!

- Criteria for Part Of Speech

  - Functional/notional tagging dominates structure in Penn Treebank, even against explicit instructions to the contrary:

Where a rule was followed: “marginal prepositions”

- Fowler (1926): “there is a continual change going on by which certain participles or adjectives acquire the character of prepositions or adverbs, no longer needing the prop of a noun to cling to”
- It is easy to have no tagging ambiguity in such cases:
  - Penn Treebank (Santorini 1991):
    - “Putative prepositions ending in -ed or -ing should be tagged as past participles (VBN) or gerunds (VBG), respectively, not as prepositions (IN),
      - According/VBG to reliable sources
      - Concerning/VBG your request of last week”

Preposition IN ⇄ Verb (VBG)

- But this makes no real sense
- Rather we see “a development caught in the act” (Fowler 1926)
- They moved slowly, toward the main gate, following the wall
- Repeat the instructions following the asterisk
- This continued most of the week following that ill-starred trip to church
- He bled profusely following circumcision
- Following a telephone call, a little earlier, Winter had said …
- IN: during [cf. endure], pending, notwithstanding
- ??: concerning, excepting, considering, …