CS 294-5: Statistical Natural Language Processing

Unlexicalized PCFGs

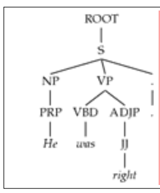
Lecture 14: 10/24/05

Treebank Sentences

```
( (S (NP-SBJ The move)
     (VP followed
         (NP (NP a round)
             (PP of
                 (NP (NP similar increases)
                     (PP by
                         (NP other lenders))
                     (PP against
                         (NP Arizona real estate loans)))))
         ,
         (S-ADV (NP-SBJ *)
                (VP reflecting
                    (NP (NP a continuing decline)
                        (PP-LOC in
                                (NP that market))))))
     .))
```

Treebank Parsing in 20 sec

- Need a PCFG for broad coverage parsing.
- Can take a grammar right off the trees (doesn't work well):



ROOT → S 1

S → NP VP . 1

NP → PRP 1

VP → VBD ADJP 1

.....

- Better results by enriching the grammar (e.g., lexicalization).
- Can also get reasonable parsers without lexicalization.

Context-Free Grammars

- A context‑free grammar is a tuple <*N, T, S, R*>
  - *N* : the set of non-terminals
    - Phrasal categories: S, NP, VP, ADJP, etc.
    - Parts-of-speech (pre-terminals): NN, JJ, DT, VB
  - *T* : the set of terminals (the words)
  - *S* : the start symbol
    - Often written as ROOT or TOP
    - *Not* usually the sentence non-terminal S
  - *R* : the set of rules
    - Of the form $X \rightarrow Y_1\ Y_2 \ldots Y_k$, with $X, Y_i \in N$
    - Examples: S → NP VP, VP → VP CC VP
    - Also called rewrites, productions, or local trees

Example CFG

- Can just write the grammar (rules with non-terminal LHSs) and lexicon (rules with pre-terminal LHSs)

Grammar

ROOT → S

S → NP VP

VP → VBP

VP → VBP NP

VP → VP PP

PP → IN NP

NP → NNS

NP → NN

NP → JJ NP

NP → NP NNS

NP → NP PP

Lexicon

JJ → new

NN → art

NNS → critics

NNS → reviews

NNS → computers

VBP → write

IN → with

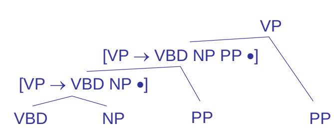
N-Ary Rules, Grammar States

- Often we want to write grammar rules like

  VP → VBD NP PP PP

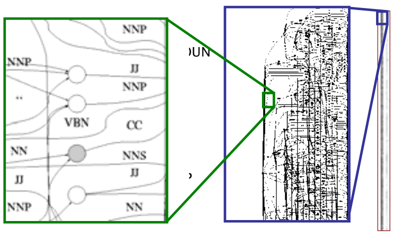
  which are not binary.

- We can work with these rules by introducing new intermediate symbols (states) into our grammar:



Treebank Grammar Scale

- Treebank grammars can be enormous!
  - As a set of FSTs, the raw grammar has ~10K states (why?).
  - Better parsers usually make the grammars larger, not smaller.

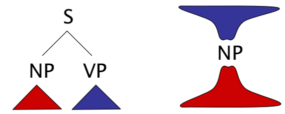


PCFGs and Independence

- Symbols in a PCFG define independence assumptions:

S $\rightarrow$ NP VP

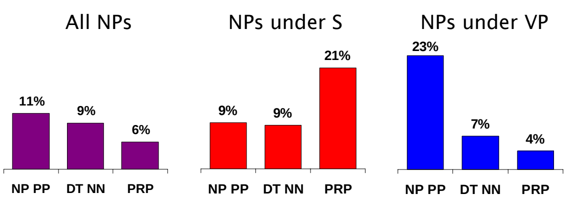
NP $\rightarrow$ DT NN



  - At any node, the material inside that node is independent of the material outside that node, given the label of that node.
  - Any information that statistically connects behavior inside and outside a node must flow through that node.

Non-Independence I

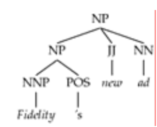
- Independence assumptions are often too strong.

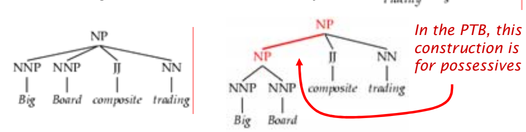


- Example: the expansion of an NP is highly dependent on the parent of the NP (i.e., subjects vs. objects).
- Also: the subject and object expansions are correlated!

Non-Independence II

- Who cares?
  - NB, HMMs, all make false assumptions!
  - For generation, consequences would be obvious.
  - For parsing, does it impact accuracy?
- Symptoms of overly strong assumptions:
  - Rewrites get used where they don't belong.
  - Rewrites get used too often or too rarely.



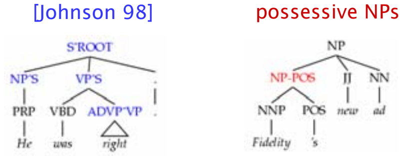


Breaking Up the Symbols

- We can relax independence assumptions by encoding dependencies into the PCFG symbols:

Parent annotation [Johnson 98]

Marking possessive NPs



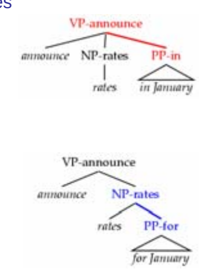
- What are the most useful "features" to encode?

Annotations

- Annotations split the grammar categories into sub-categories (in the original sense).
- Conditioning on history vs. annotating
  - P(NP^S $\rightarrow$ PRP) is a lot like P(NP $\rightarrow$ PRP | S)
  - P(NP-POS $\rightarrow$ NNP POS) isn't history conditioning.
- Feature / unification grammars vs. annotation
  - Can think of a symbol like NP^NP-POS as NP [parent:NP, +POS]
- After parsing with an annotated grammar, the annotations are then stripped for evaluation.

Lexicalization

- Lexical heads important for certain classes of ambiguities (e.g., PP attachment):
- Lexicalizing grammar creates a much larger grammar. (cf. next week)
  - Sophisticated smoothing needed
  - Smarter parsing algorithms
  - More data needed
- How necessary is lexicalization?
  - Bilexical vs. monolexical selection
  - Closed vs. open class lexicalization



Unlexicalized PCFGs

- What is meant by an "unlexicalized" PCFG?
  - Grammar not systematically specified to the level of lexical items
    - NP [stocks] is not allowed
    - NP^S-CC is fine
  - Closed vs. open class words (NP^S [the])
    - Long tradition in linguistics of using function words as features or markers for selection
    - Contrary to the bilexical idea of semantic heads
    - Open-class selection really a proxy for semantics
- Honesty checks:
  - Number of symbols: keep the grammar very small
  - No smoothing: over-annotating is a real danger
  - No smoothing is a bad idea – this use is rhetorical!

Typical Experimental Setup

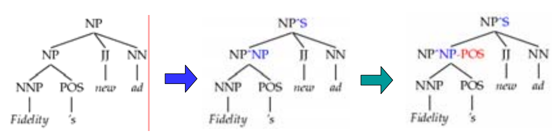
- Corpus: Penn Treebank, WSJ

Training:	sections	02-21
Development:	section	22 (here, first 20 files)
Test:	section	23

- Accuracy – F1: harmonic mean of per-node labeled precision and recall.
- Here: also size – number of symbols in grammar.
  - Passive / complete symbols: NP, NP^S
  - Active / incomplete symbols: NP → NP CC •

Multiple Annotations

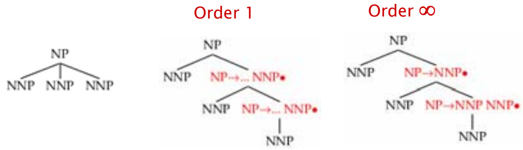
- Each annotation done in succession
  - Order does matter
  - Too much annotation and we'll have sparsity issues (where?).

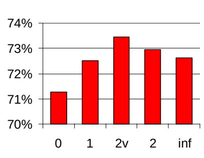


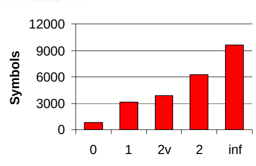
Horizontal Markovization

Order 1

Order ∞







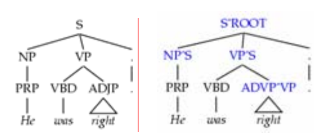
Vertical Markovization

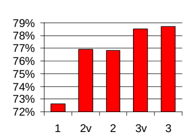
- Vertical Markov order: rewrites depend on past *k* ancestor nodes.

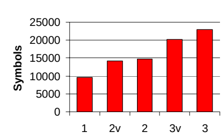
  (cf. parent annotation)

Order 1

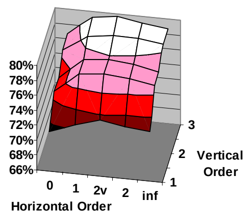
Order 2

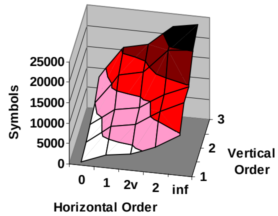






Vertical and Horizontal



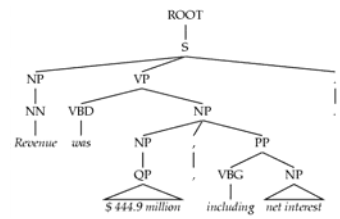


- Examples:
  - Raw treebank: v=1, h=∞
  - Johnson 98: v=2, h=∞
  - Collins 99: v=2, h=2
  - Best F1: v=3, h=2v

Model	F1	Size
Base: v=h=2v	77.8	7.5K

Unary Splits

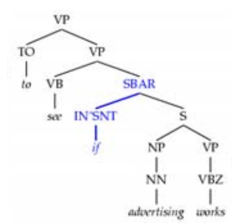
- Problem: unary rewrites used to transmute categories so a high-probability rule can be used.
- Solution: Mark unary rewrite sites with -U



Annotation	F1	Size
Base	77.8	7.5K
UNARY	78.3	8.0K

Tag Splits

- Problem: Treebank tags are too coarse.
- Example: Sentential, PP, and other prepositions are all marked IN.
- Partial Solution:
  - Subdivide the IN tag.



Annotation	F1	Size
Previous	78.3	8.0K
SPLIT-IN	80.3	8.1K

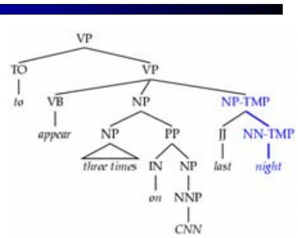
Other Tag Splits

- UNARY-DT: mark demonstratives as DT^U ("the X" vs. "those")
- UNARY-RB: mark phrasal adverbs as RB^U ("quickly" vs. "very")
- TAG-PA: mark tags with non-canonical parents ("not" is an RB^VP)
- SPLIT-AUX: mark auxiliary verbs with –AUX [cf. Charniak 97]
- SPLIT-CC: separate "but" and "&" from other conjunctions
- SPLIT-%: "%" gets its own tag.

F1	Size
80.4	8.1K
80.5	8.1K
81.2	8.5K
81.6	9.0K
81.7	9.1K
81.8	9.3K

Treebank Splits

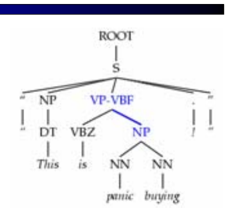
- The treebank comes with some annotations (e.g., -LOC, -SUBJ, etc).
  - Whole set together hurt the baseline.
  - One in particular is very useful (NP-TMP) when pushed down to the head tag (why?).
  - Can mark gapped S nodes as well.



Annotation	F1	Size
Previous	81.8	9.3K
NP-TMP	82.2	9.6K
GAPPED-S	82.3	9.7K

Yield Splits

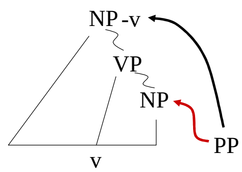
- Problem: sometimes the behavior of a category depends on something inside its future yield.
- Examples:
  - Possessive NPs
  - Finite vs. infinite VPs
  - Lexical heads!
- Solution: annotate future elements into nodes.
  - Lexicalized grammars do this (in very careful ways – why?).



Annotation	F1	Size
Previous	82.3	9.7K
POSS-NP	83.1	9.8K
SPLIT-VP	85.7	10.5K

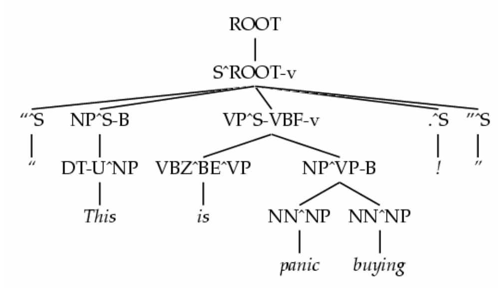
Distance / Recursion Splits

- Problem: vanilla PCFGs cannot distinguish attachment heights.
- Solution: mark a property of higher or lower sites:
  - Contains a verb.
  - Is (non)-recursive.
    - Base NPs [cf. Collins 99]
    - Right-recursive NPs



Annotation	F1	Size
Previous	85.7	10.5K
BASE-NP	86.0	11.7K
DOMINATES-V	86.9	14.1K
RIGHT-REC-NP	87.0	15.2K

A Fully Annotated (Unlex) Tree



Some Test Set Results

Parser	LP	LR	**F1**	CB	0 CB
Magerman 95	84.9	84.6	**84.7**	1.26	56.6
Collins 96	86.3	85.8	**86.0**	1.14	59.9
Unlexicalized	86.9	85.7	**86.3**	1.10	60.3
Charniak 97	87.4	87.5	**87.4**	1.00	62.1
Collins 99	88.7	88.6	**88.6**	0.90	67.1

- Beats "first generation" lexicalized parsers.
- Lots of room to improve – more complex models next.