So far ...

1. PTB WSJ
2. BIKEL COLLINS
3. EVALB
Bikel Collins and EVALB

- Performance of Bikel-Collins on section 23

<table>
<thead>
<tr>
<th>Model</th>
<th>Performance on Section 00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR</td>
</tr>
<tr>
<td>Collins’ Model 2</td>
<td>89.75</td>
</tr>
<tr>
<td>Baseline (Model 2 emulation)</td>
<td>89.89</td>
</tr>
<tr>
<td>Clean-room Model 2</td>
<td>88.85</td>
</tr>
</tbody>
</table>

Figure 15
Overall parsing results using only details found in (Collins, 1997; Collins, 1999). The first two lines show the results of Collins’ parser and those of our parser in its “complete” emulation mode (i.e., including unpublished details). All reported scores are for sentences of length ≤ 40 words. LR/LP are the primary scoring metrics, labeled precision and labeled recall, respectively. CBs is the number of crossing brackets. 0 CBs and ≤ 2 CBs are the percentage of sentences with 0 and ≤ 2 crossing brackets, respectively. F (the “F-measure”) is the evenly-weighted harmonic mean of precision and recall, or \( \frac{2 \cdot LP \cdot LR}{LP + LR} \).
System Flowchart

• Diagram:

- **WSJ treebank 00–24**
  - tregex View Search
    - Treebank trees .mrg
      - create using cat
    - Treebank sentences .txt
      - How?
    - Sec 23 trees .mrg
      - create using cat
    - tregex tsgurone (-s flag)
      - Sec 23 gold trees .mrg
        - (one tree per line)
      - COLLINS .prm
  - Bikel Collins parser Train
    - Events .obj.gz
    - Parse trees .txt.parsed
      - (one tree per line)
  - EVALB
    - recall precision F-measure ≈86%
Testing

• Ungraded homework exercise:
  • Run Bikel-Collins on Section 23

Penn Treebank (WSJ)

training
sections 2–21

almost 40K sentences (39,832)

2.4K sentences (2,416)

section 23

evaluation

0

24
Section 23 sentences

- Need to run the Bike-Collins parser on Section 23 sentences:
  - How to extract them?
  - Several possible ways... use (1) NLTK or (2) TREEBANK files directly or (3) tregex
Section 23 sentences

• Use the POS tagged data in TREEBANK_3:
Section 23 sentences

• Use the POS tagged data in TREEBANK_3:

```
2
3
4==================================
5
6No/RB ,/
7[ it/PRP ]
8
9[ was/VBD n't/RB Black/NNP Monday/NNP ]
10./.
11
12==================================
13
14But/CC while/IN
15[ the/DT New/NNP York/NNP Stock/NNP Exchange/NNP ]
16did/VBD n't/RB
17[ fall/VB ]
18apart/RB
19[ Friday/NNP ]
20as/IN
21[ the/DT Dow/NNP Jones/NNP Industrial/NNP Average/NNP ]
22plunged/VBD
23[ 190.58/CD points/NNS ]
```
Section 23 sentences

- Use the section trees: and write a regex:
Extracting the sentences

• Use tregex Save matched sentences... :
Extracting the sentences

- Then delete prefix, and empty elements (see patterns)

```plaintext
wsj_2300.mrg-1 No, it wasn’t Black Monday.
swj_2300.mrg-2 But while the New York Stock Exchange did n’t fall apart Friday as the Dow Jones Industrial Average plunged 190.58 points — most of it in the final hour — it barely managed *-2 to stay this side of chaos.
swj_2300.mrg-3 Some ‘‘ circuit breakers ’’ installed after the October 1987 crash failed their first test, traders say 0 *T*-1, *-2 unable *-3 to cool the selling panic in both stocks and futures.
swj_2300.mrg-4 The 49 stock specialist firms on the Big Board floor — the buyers and sellers of last resort who *T*-2 were criticized *-1 after the 1987 crash — once again could n’t handle the selling pressure.
swj_2300.mrg-5 Big investment banks refused *-1 to step up to the plate *-2 to support the beleaguered floor traders by *-3 buying big blocks of stock, traders say 0 *T*-4.
swj_2300.mrg-6 Heavy selling of Standard & Poor’s 500-stock index futures in Chicago relentlessly beat stocks downward.
swj_2300.mrg-7 Seven Big Board stocks — UAL, AMR, BankAmerica, Walt Disney, Capital Cities/ABC, Philip Morris and Pacific Telesis Group — stopped *-1 trading and never resumed.
swj_2300.mrg-8 The finger-pointing has already begun.
swj_2300.mrg-9 ‘‘The equity market was illiquid."
wsj_2300.mrg-10 Once again — LCB — the specialists — RCB — were not able *-3 to handle the imbalances on the floor of the New York Stock Exchange, ” said *T*-2 Christopher Pedersen, senior vice president at Twenty-First Securities Corp.
wsj_2300.mrg-11 Countered 0 *ICH*-2 James Maguire, chairman of specialists Henderson Brothers Inc.: ‘‘ It *EXP*-1 is easy ” to say 0 the specialist is n’t doing his job.
wsj_2300.mrg-12 When the dollar is in a free-fall *T*-1, even central banks can n’t stop it.
wsj_2300.mrg-13 Speculators are calling for a degree of liquidity that *T*-1 is not there in the market. ”
wsj_2300.mrg-14 Many money managers and some traders had already left their offices early Friday afternoon on a warm autumn day — because the stock market was so quiet.
wsj_2300.mrg-15 Then in a lightning plunge, the Dow Jones industrials in barely an hour surrendered about a third of their gains this year, *-1 chalking up a 190.58-point, or 6.9 %, loss on the day in gargantuan trading volume.
```
Training: Robustness and Sensitivity

• (Bikel 2004):
  • “it may come as a surprise that the [parser] needs to access more than 219 million probabilities during the course of parsing the 1,917 sentences of Section 00 [of the PTB]."

From lecture 11
Training: Robustness and Sensitivity

• Trainer has a memory like a phone book:

76.8% singular events
94.2% 5 or fewer occurrences
Homework 7

• Let's explore the question: how much data do we need to get good parsing performance?

• WSJ PTB: there exist sections 00 through 24
• Using .mrg files from an increasing number of randomly selected sections (excluding section 23)
• Train the Bikel-Collins parser
• Test the trained Bikel-Collins parser on Section 23 (WARNING: SLOW!)
• Run EVALB to compare output of Bikel-Collins against gold-standard trees
• Plot a graph of the F-score
Homework 7

• **Training**: normally 02-21 (20 sections)
  • concatenate mrg files as input to Bikel Collins trainer
  • training is **fast**
• **Evaluation**: on section 23
  • use tsurgeon to get tree data ready for EVALB (1 tree per line)
  • Use nltk or the .pos or .mrg files or tregex to grab the sentences for parsing
  • 2400+ sentences in section 23, parsing is **slow**…
  • Can split the sentences up and run on multiple processors
Training: Robustness and Sensitivity

.observed file
Frequency 1 observed data for:
(NP (NP (DT a) (NN milk))(PP (IN with)(NP (ADJP (CD 4) (NN %))(NN butterfat))))

• (mod ((with IN) (milk NN) PP (+START+) ((+START+ +START+)) NP-A NPB () false right) 1.0)
  • modHeadWord (with IN)
  • headWord (milk NN)
  • modifier PP
  • previousMods (+START+)
  • previousWords ((+START+ +START+))
  • parent NP-A
  • head NPB
  • subcat ()
  • verbIntervening false
  • side right

• (mod ((+STOP+ +STOP+) (milk NN) +STOP+ (PP) ((with IN))) NP-A NPB () false right) 1.0)
  • modHeadWord (+STOP+ +STOP+)
  • headWord (milk NN)
  • modifier +STOP+
  • previousMods (PP)
  • previousWords ((with IN))
  • parent NP-A
  • head NPB
  • subcat ()
  • verbIntervening false
  • side right

More on this next lecture ...
Generating Training Data

• One possible way (*old Perl code of mine*):

```perl
use list::Util 'shuffle';
my $prefix = '/Users/sandiway/research/';
my $dbprefix = $prefix/dbparser;
my @sections = ('02', '03', '04', '05', '06', '07', '08', '09',
                '10', '11', '12', '13', '14', '15', '16', '17', '18',
                '19', '20', '21);
# from http://perldoc.perl.org/perlfaq4.html
my @shuffled = shuffle(@sections);

sub trainparser {  
    # assume touch wsj-.mrg (i.e. empty file)
    my $filename = "wsj-";
    while (@shuffled) {  
        my $section = shift @shuffled;
        system "cat $filename.mrg wsj_$section.mrg > $filename$section.mrg";
        print "Training on $filename$section.mrg\n";
        system "$dbprefix/bin/train 1000 $dbprefix/settings/collins.properties $filename$section.mrg";
        $filename .= "$section;"
    }
}
trainparser;
```

function `shuffle` permutes `@sections`
Training Data

• What does the graph for the F-score (Labeled Recall/Labeled Precision) look like?

![Graph of F-score vs. # of sentences used for training]

*report your results*
Homework 7

• Submit on March 12th (Monday by midnight) your graph
• Write up:
  • How you extracted the sentences in section 23 for parsing
  • How you selected an increasing number of .mrg files for training
• One PDF file
• Don't include your training files!