Bikel Collins and EVALB

- Did anyone try with Bikel-Collins on section 23?
Steps

• Running evalb

• Need:
  – parameter file: COLLINS.prm
  – Parser output: test.txt.parsed
  – Gold file: test.gold
Steps

- (Assume my data files are on my Desktop and EVALB is installed in my research subdirectory)
  - `../research/EVALB/evalb -p ../research/EVALB/COLLINS.prm test.gold test.txt.parsed`

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>83.33</td>
<td>100.00</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>83.33</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>84.38</td>
<td>90.00</td>
<td>27</td>
<td>32</td>
<td>30</td>
<td>1</td>
<td>37</td>
<td>37</td>
<td>100.00</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>68.18</td>
<td>75.00</td>
<td>15</td>
<td>22</td>
<td>20</td>
<td>2</td>
<td>26</td>
<td>26</td>
<td>100.00</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>90.91</td>
<td>90.91</td>
<td>20</td>
<td>22</td>
<td>22</td>
<td>1</td>
<td>32</td>
<td>32</td>
<td>100.00</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>91.67</td>
<td>95.65</td>
<td>22</td>
<td>24</td>
<td>23</td>
<td>0</td>
<td>24</td>
<td>23</td>
<td>95.83</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>84.62</td>
<td>91.67</td>
<td>11</td>
<td>13</td>
<td>12</td>
<td>0</td>
<td>16</td>
<td>16</td>
<td>100.00</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>83.33</td>
<td>83.33</td>
<td>15</td>
<td>18</td>
<td>18</td>
<td>1</td>
<td>22</td>
<td>21</td>
<td>95.45</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>83.33</td>
<td>100.00</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>80.00</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>80.00</td>
<td>100.00</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>100.00</td>
</tr>
</tbody>
</table>

--- Summary ---

- All
  - Number of sentence = 9
  - Number of Error sentence = 0
  - Number of Skip sentence = 0
  - Number of Valid sentence = 9
  - Bracketing Recall = 83.78
  - Bracketing Precision = 89.21
  - Bracketing FMeasure = 86.41
  - Complete match = 0.00
  - Average crossing = 0.56
  - No crossing = 55.56
  - 2 or less crossing = 100.00
  - Tagging accuracy = 97.69

Recall 83.8, Precision 89.2
Task from last time

• On website (if you have trouble):
  - tregex_wsj_23.txt

1 No, it was n't Black Monday.
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 to stay this side of chaos.
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

- **Bikel Collins parser**
  - **Train**

- **Sec 23 gold trees .mrg**
  - (one tree per line)

- **tregex tsurgeon (-s flag)**

- **Events .obj.gz**
- **Parse trees .txt.parsed**
  - (one tree per line)
- **EVALB**
  - recall precision
  - F-measure ≈86%
  - COLLINS .prm
Extracting the sentences

We need to extract the sentences to run the Bikel Collins parser

• Section 23 contains 2416 sentences

• There are many possible ways to do this: let's look at a couple ...
Extracting the sentences

- Use the sec 23 gold trees: Perl regex:
Extracting the sentences

- Use the POS tagged data in TREEBANK_3:
Extracting the sentences

- Use the POS tagged data in TREEBANK_3:

```
No/RB ,/, [ it/PRP ]
[ was/VBD n't/RB Black/NNP Monday/NNP ]
./.

But/CC while/IN
[ the/DT New/NNP York/NNP Stock/NNP Exchange/NNP ]
did/VBD n't/RB
[ fall/VB ]
apart/RB
[ Friday/NNP ]
as/IN
[ the/DT Dow/NNP Jones/NNP Industrial/NNP Average/NNP ]
plunged/VBD
[ 190.58/CD points/NNS ]
```
Extracting the sentences

• Use tregex Save matched sentences:

From file: /Users/sandiway/research/TREEBANK_3/parsed/mrg/wsj/23/wsj_2399.mrg
Extracting the sentences

• Use tregex Save Sentences: Perl regex

```
\*T\*-\-[0-9]+\n\*-\-[0-9]+\n\*U\*
\b0\b
\b\*\b
```

---

```
wsj_2300.mrg-1  No, it was n't Black Monday.
wsj_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.
wsj_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.
wsj_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.
wsj_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.
wsj_2300.mrg-6  Heavy selling of Standard & Poor's 500-stock index futures in Chicago relented downward.
wsj_2300.mrg-7  Seven Big Board stocks -- UAL, AMR, BankAmerica, Walt Disney, Capital Cities/ABC and Pacific Telesis Group -- stopped *-1 trading and never resumed.
wsj_2300.mrg-8  The finger-pointing has already begun.
wsj_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 imploding New York Stock Exchange, said *T*-2 Christopher Pedersen, senior vice president at Two.
wsj_2300.mrg-11  Countered 0 *ICH*-2 James Maguire, chairman of specialists Henderson Bros.
wsj_2300.mrg-12  When the dollar is in a free-fall *T*-1, even central banks can'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.
```
Summary

- **WSJ corpus**: sections 00 through 24
- **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 data ready for EVALB
  - Use either the .pos 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
New Homework

• **Question:**
  – How much training data do you need?

• **Homework:**
  – How does Bikel Collins vary in precision and recall on test data if you randomly pick 1..24 out of 25 sections to do the training with?

• **Test section:**
  – I want you to pick a test section that’s not section 23

• **Use EVALB**
  – plot precision and recall graphs
Summary

• **Question:**
  – How much training data do you need?

• **Homework:**
  – How does Bikel Collins vary in precision and recall on test data if you randomly pick 1..24 out of 25 sections to do the training with?

• **Test section:**
  – I want you to pick a test section that’s not section 23

• **Use EVALB**
  – plot precision and recall graphs
Perl Code

• Generate training data

```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');

my @shuffled = shuffle(@sections);

sub trainparser {
  # assume touch wsj-1.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;
```
<table>
<thead>
<tr>
<th>File Name</th>
<th>Size</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>wsj-120520211011091716180806030714.mrg</td>
<td>20.1 MB</td>
<td>Today 5:58 PM</td>
</tr>
<tr>
<td>wsj-1205202110110917161808060307.observed.gz</td>
<td>7.4 MB</td>
<td>Today 5:57 PM</td>
</tr>
<tr>
<td>wsj-12052021101109171618080603.mrg</td>
<td>18.6 MB</td>
<td>Today 5:57 PM</td>
</tr>
<tr>
<td>wsj-12052021101109171618080603.obj.gz</td>
<td>14.9 MB</td>
<td>Today 5:57 PM</td>
</tr>
<tr>
<td>wsj-12052021101109171618080603.observed.gz</td>
<td>6.8 MB</td>
<td>Today 5:57 PM</td>
</tr>
<tr>
<td>wsj-12052021101109171618080603.mrg</td>
<td>19.1 MB</td>
<td>Today 5:56 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>14.3 MB</td>
<td>Today 5:56 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.obj.gz</td>
<td>6.5 MB</td>
<td>Today 5:56 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>13.3 MB</td>
<td>Today 5:56 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.mrg</td>
<td>16.1 MB</td>
<td>Today 5:56 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>6 MB</td>
<td>Today 5:55 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.mrg</td>
<td>14.8 MB</td>
<td>Today 5:55 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.obj.gz</td>
<td>13.1 MB</td>
<td>Today 5:55 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>5.9 MB</td>
<td>Today 5:55 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.mrg</td>
<td>14.4 MB</td>
<td>Today 5:54 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.obj.gz</td>
<td>12 MB</td>
<td>Today 5:54 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>5.4 MB</td>
<td>Today 5:54 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.mrg</td>
<td>12.9 MB</td>
<td>Today 5:54 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.obj.gz</td>
<td>10.5 MB</td>
<td>Today 5:54 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>4.6 MB</td>
<td>Today 5:53 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.obj.gz</td>
<td>9.7 MB</td>
<td>Today 5:53 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.mrg</td>
<td>11 MB</td>
<td>Today 5:53 PM</td>
</tr>
<tr>
<td>wsj-120520211011091716180806.observed.gz</td>
<td>4.2 MB</td>
<td>Today 5:53 PM</td>
</tr>
</tbody>
</table>
Training Data

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

F-score

# of sentences used for training

report your results next time...
Robustness and Sensitivity

it’s often assumed that statistical models are less brittle than symbolic models

1. can get parses for ungrammatical data
2. are they sensitive to noise or small perturbations?
Robustness and Sensitivity

Examples

1. Herman mixed the water with the milk
2. Herman mixed the milk with the water
3. Herman drank the water with the milk
4. Herman drank the milk with the water

\[ f(\text{water}) = 117, \quad f(\text{milk}) = 21 \]
Robustness and Sensitivity

Examples

1. Herman mixed the water with the milk
2. Herman mixed the milk with the water
3. Herman drank the water with the milk
4. Herman drank the milk with the water

Different PP attachment choices

logprob = -50.4

logprob = -47.2
Robustness and Sensitivity

First thoughts...
• does milk forces low attachment?
  (high attachment for other nouns like water, toys, etc.)
  *Is there something special about the lexical item milk?*
• 24 sentences in the WSJ Penn Treebank with milk in it, 21 as a noun
First thoughts...

Is there something special about the lexical item milk?

• 24 sentences in the WSJ Penn Treebank with milk in it, 21 as a noun
• but just one sentence (#5212) with PP attachment for milk

Could just one sentence out of 39,832 training examples affect the attachment options?
Robustness and Sensitivity

- Simple perturbation experiment
  - alter that one sentence and retrain
Robustness and Sensitivity

- Simple perturbation experiment
  - alter that one sentence and retrain

delete the PP with 4% butterfat altogether
Robustness and Sensitivity

- **Simple perturbation experiment**
  - alter that one sentence and retrain

The Bikel/Collins parser can be retrained quickly by altering one sentence and retraining. Alternatively, one could bump it up to the VP level.
Robustness and Sensitivity

- **Result:**
  - high attachment for previous PP adjunct to *milk*

Could just one sentence out of 39,832 training examples affect the attachment options?

Why such extreme sensitivity to perturbation? Logprobs are conditioned on many things; hence, lots of probabilities to estimate

- **smoothing**
- need every piece of data, even low frequency ones
Details...

- Two sets of files:

```
dhcp-10-134-211-43:bin sandiway$ diff wsj-02-21-milk.mrg wsj-02-21.mrg
31530,31532c31530,31537
<     (NP (DT a) (NN milk))
<     (PP-LOC (IN in)
<       (NP (DT the) (NNP South)))
---
>     (NP
>     (NP (DT a) (NN milk))
>     (PP (IN with)
>       (NP
>         (ADJP (CD 4) (NN %))
>         (NN butterfat)))
>     (PP-LOC (IN in)
>       (NP (DT the) (NNP South)))
```
Bikel/Collins Parser wrapper

Herman drank the milk with the water.
Herman drank the water with the milk.

MXPOST:

((Herman (NNP)) (drank (VB)) (the (DT)) (milk (NN)) (with (IN)) (the (DT)) (water (NN)))
((Herman (NNP)) (drank (VB)) (the (DT)) (water (NN)) (with (IN)) (the (DT)) (milk (NN)))

Bikel/Collins: Parse
Train

using Derived Counts: wsj-02-21-milk.obj.gz
using Treebank: wsj-02-21.mrg
using settings: collins.properties

Treebankviewer: Display

Details...
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]."
Robustness and Sensitivity

• Trainer has a memory like a phone book:
Robustness and Sensitivity

Frequency 1 observed data for:

\[\text{NP (NP (DT a)(NN milk))(PP (IN with)(NP (ADJP (CD 4)(NN %))(NN butterfat))))}\]

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

- \((\text{mod ((+STOP+ +STOP+) (milk NN) +STOP+ (PP) ((with IN)) NP-A NPB () false right) 1.0})\)
  - modHeadWord (+STOP+ +STOP+)
  - headWord (milk NN)
  - modifier +STOP+
  - previousMods (PP)
  - previousWords ((with IN))
  - parent NP-A
  - head NPB
  - subcat ()
  - verbIntervening false
  - side right
Robustness and Sensitivity

B-Collins: Observed Data

76.8% singular events
94.2% 5 or fewer occurrences
Robustness and Sensitivity

• Full story more complicated than described here...

• by picking different combinations of verbs and nouns, you can get a range of behaviors

<table>
<thead>
<tr>
<th>Verb</th>
<th>Noun</th>
<th>Attachment</th>
</tr>
</thead>
<tbody>
<tr>
<td>drank</td>
<td>water</td>
<td>high</td>
</tr>
<tr>
<td>mixed</td>
<td>water</td>
<td>low</td>
</tr>
<tr>
<td>mixed</td>
<td>computer</td>
<td>low</td>
</tr>
</tbody>
</table>