LING/C SC 581:
Advanced Computational Linguistics

Lecture Notes
Jan 27th
Today's Topics

• Homework 1 Review
• Penn Treebank (V3)
• Homework 2: Install Stanford's tregex
Homework 1 Review

• Number of states

LR(0)
5 % State 0
6 state(0, 'np->.np pp').
7 state(0, 'np->.nnp').
8 state(0, 'np->.dt nn').
9 state(0, 's->.np vp').
10 state(0, 'ss->.s $').
89
90 % State 14
91 state(14, 'ss->s $.').
92 goTo(1, $, 14).

LR(1)
5 % State 0
6 state(0, 'np->.np pp/[in,vbd,vbz]').
7 state(0, 'np->.nnp/[in,vbd,vbz]').
8 state(0, 'np->.dt nn/[in,vbd,vbz]').
9 state(0, 's->.np vp/$]').
10 state(0, 'ss->.s $/[[]]').
119 % State 20
120 state(20, 'ss->s $./[[]]').
121 goTo(1, $, 20).
Homework 1 Review

• LR(0) Machine
• Consider state 12:
  – shift IN or reduce PP
  – not deterministic!
  – therefore not LR(0)
Homework 1 Review

• LR(1) Machine
• Consider state 18:
  – lookahead sets are the same POS tags
  – shift IN or reduce PP
  – not LR(1)!
Homework 1 Review

• grammar0.pl is neither LR(0) nor LR(1)
Homework 1 Review

- *John saw the boy with the telescope*
- LR(0)
Homework 1 Review

- John saw the boy with the telescope
- LR(0)

```plaintext
Accepting state
Input: []
Stack:
[s(npp(john)), vp(vbd(saw), np(dt(the), nn(boy))), pp(in(with), np(dt(the), nn(telescope)))]
State stack: accept

P = [s(nnp(john)), vp(vbd(saw), np(dt(the), nn(boy))), pp(in(with), np(dt(the), nn(telescope)))]
State 1:
ss->s .$
Input: [in(with), dt(the), nn(telescope), $]
Stack: [s(nnp(john)), vp(vbd(saw), np(dt(the), nn(boy)))]
State stack: [1,0]
false.

egrep 'State \d+|Accepting' q4lr0.txt | wc -l
26
```
Homework 1 Review

- *John saw the boy with the telescope*
- LR(1)

```bash
egrep 'State \d+|Accepting' q4lr1.txt | wc -l
```

25
Homework 1 Review

- *John saw the boy with a limp with the telescope*
- 5 parses:

  ![Parse Tree 1]
  ![Parse Tree 2]
Homework 1 Review

• *John saw the boy with a limp with the telescope*

• 5 parses:
Homework 1 Review

- *John saw the boy with a limp with the telescope*
- 5 parses:
Homework 1 Review

• John saw the boy with a limp with the telescope

• 5 parses:

- see, boy, limp, telescope
- see, boy, telescope, limp
- see, boy, limp, telescope
- see, telescope, boy, limp
- see, telescope, boy, limp
Homework 1 Review

- **Number of states visited**
  - $\$ egrep 'State \d+|Accepting' q5lr0.txt | wc -l
  - 70
  - $ egrep 'State \d+|Accepting' q5lr1.txt | wc -l
  - 67
Homework 1 Review

- Merge states?
  - 15: reduce NP (|RHS| = 2)
  - 16: reduce NP (|RHS| = 2)
  - 15: Lookahead \{IN, $\}
  - 16: Lookahead \{IN, $\}
Context-free Grammars

• Treebanks
  – (538: Context-free grammars)

• regex search over parse trees
  – (538: Perl regex on strings)
Homework 2: Install Tregex


Computer language: java
Homework 2: Install Tregex

- We’ll use the program tregex to explore the Penn Treebank
  – current version:

  Download

  Download Tregex version 3.5.2 (source and executables for all platforms)
  Download Tregex version 3.4 Mac OS X disk image (GUI packaged as Mac application; Java 1.7 runtime included)
Penn Treebank

• Availability
  – Source:
    • Linguistic Data Consortium (LDC)
    • U. of Arizona is a (fee-paying) member of this consortium
    • Resources are made available to the community through the main library
  • URL
    – http://sabio.library.arizona.edu/search/X
Penn Treebank (V3)

<table>
<thead>
<tr>
<th>Location</th>
<th>Main Library</th>
<th>Call #</th>
<th>P128.D37 T74 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>The Penn Treebank project, release 3 [computer file]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>CALL NO.</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Library</td>
<td>P128.D37 T74 1999</td>
<td>IN LIBRARY</td>
</tr>
</tbody>
</table>

- Call Record

Expanded Call Record:

Other title: Title in LDC documentation: *Treebank-3 Wall Street journal*.

Other auth: *Linguistic Data Consortium*.

Publisher: *[Philadelphia, Pa.]: Linguistic Data Consortium, [1999]*.


Summary: The corpus consists of "1 million words of 1989 Wall Street Journal material annotated in *Treebank II* style, a small sample of ATIS-3 material annotated in *Treebank II* style, a fully tagged version of the Brown corpus".

Note: Title from disc label.

Description: "LDC99T42" -- on-line documentation.


1585631639

Have it on a usb drive here ...
Penn Treebank (V3)

- Raw data:
Penn Treebank

1. Tagging Guide
2. Arpa94 paper
3. Parse Guide
Penn Treebank
Penn Treebank

sections 00-24
Penn Treebank
tregex

- Tregex is a Tgrep2-style utility for matching patterns in trees.

```bash
#!/bin/sh
java -mx500m -cp `dirname $0`/stanford-tregex.jar edu.stanford.nlp.trees.tregex.gui.TregexGUI
```

- `mx` flag, the 300m default memory size may need to be increased depending on the platform
tregex

• Select the PTB directory
  – TREEBANK_3/parsed/mrg/wsj/ Deselect any unwanted files

• Browse
tregex

- **Search**
  - NP-SBJ << (dominates) vs. < (immediately dominates) NNP

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
tregex

A < B
  A immediately dominates B
A > B
  A is immediately dominated by B
A $ B
  A is a sister of B (and not equal to B)
A .. B
  A precedes B
A . B
  A immediately precedes B
A ,, B
  A follows B
A , B
  A immediately follows B
A <<= B
  B is a leftmost descendent of A
A <<= B
  B is a rightmost descendent of A