Practical CCG

CCGBank
Tgrep2
CCG Parse
CCGBank

Julia Hockenmaier and Mark Steedman

CCGBank: translation of Penn Tree Bank to CCG

Pairs syntactic derivations with word-word dependencies

CCGBank covers 99.44% of Penn Tree Bank (and fixes some errors/inconsistencies)

Supports Tgrep access
### Example Lexical Entries

**Entry**

**Word** (mail)

**Category** (N)

**Word probability** ($P(\text{mail} \mid N)$)

**Category probability** ($P(N \mid \text{mail})$)

**Word-category frequency in DB**

<table>
<thead>
<tr>
<th>mail</th>
<th>N</th>
<th>0.00013495</th>
<th>0.508772</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>mail</td>
<td>N/N</td>
<td>0.000168503</td>
<td>0.438596</td>
<td>25</td>
</tr>
<tr>
<td>mail</td>
<td>S[b]\NP</td>
<td>0.000367647</td>
<td>0.0175439</td>
<td>1</td>
</tr>
<tr>
<td>mail</td>
<td>((S[b]\NP)/NP)/NP</td>
<td>0.00359712</td>
<td>0.0175439</td>
<td>1</td>
</tr>
<tr>
<td>mail</td>
<td>((S[dc]\NP)/PP)/NP</td>
<td>0.000619963</td>
<td>0.0175439</td>
<td>1</td>
</tr>
</tbody>
</table>
CCGBank access

Within CMU network
/afs/cs.cmu.edu/academic/class/11722-s08/DATA/CCG

Manual
  142 page

Corpus
  2713 paragraphs
  48,934 sentences
  83,540 lexical entries
Basic UPenn tree

```
ADVP:a
  |    
  just

VBZ:h
  |    
  opened

NP:c
  |    
  its doors

PP-TMP:a
  |    
  in July
```

Tree Conversion
Converted to Binary Branching Tree

```
VP
 /     \           /     \       /     \
ADVP:a  VP:h      PP-TMP:a   VP:h
       |           |           |
       just       VP:h       NP:c
       |           |           |
       opened    in July  in July
       |           |           |
       its doors |           |
```
Re-labelled CCG tree

\[
\begin{array}{c}
\text{S[dcl]\ NP} \\
\text{(S\ NP)/(S\ NP) S[dcl]\ NP} \\
\text{just \textbf{S[dcl]\ NP \ (S\ NP)\ (S\ NP) \ NP}} \\
\text{(S[dcl]\ NP)/NP \ NP in July}} \\
\text{opened its doors}
\end{array}
\]
• Sentences identified by type:
  • $S[dcl]$, $S[pss]$, $S[q]$, $S[b]$
  • $S[pss] \setminus NP$ (passive VP)
Example parse

\[
\begin{align*}
\text{Did} & \quad \text{I} & \quad \text{buy} & \quad \text{it?} \\
S[q]/(S[b]\smallsetminus NP)/NP & \quad \overrightarrow{NP} & \quad (S[b]\smallsetminus NP)/NP) & \quad \overrightarrow{NP} \\
S[q]/(S[b]\smallsetminus NP) & \quad \overrightarrow{NP} & \quad S[b]\smallsetminus NP & \quad \overrightarrow{NP} \\
S[q] & \quad \overrightarrow{NP}
\end{align*}
\]
• Penn TreeBank, compounds are flat
  Conversions must build binary branching

(NP (DT this)
  (NN consumer)
  (NNS electronics)
  (CC and)
  (NNS appliances)
  (NN retailing)
  (NN chain))
• Give an engraving to Chapman
• Give to Chapman an engraving by Rembrandt
• Shifts aren’t identified in TreeBank

```
Actual Heavy NP Treatment

The surge
   
   ((S[dcl]\NP)/NP)/PP  PP  the number of country funds ... 
                  |      
               brings  to nearly 50

(a declarative sentence) The surge brings the number of country funds to nearly 50.
```
Tgrep2 searching

- ./tgrep2 –c data/ccgbank.00-24.t2c EXPR
  - Where EXPR may be
  - ‘S\[pss\]’</NP//S[\pss\]/’
  - “/.*/<^[Tt]hat/”

- Regex notes:
  - **Backslash the backslash**
  - **Backslash [ ] literals**
  - **Practice a lot …**
Some specific examples

Page 29 of manual
Some problems

Due to original syntactic derivations
compound noun structure left as flat
Co-indexing issues
“I was early yesterday and late today”
early/late modify was, should be separate
events
Fails to capture control dependencies in
argument cluster coordination
“I want him to go and you to stay.”
Translation Algorithm

Translate each parse tree to word-word dependencies

For each tree:
   determineConstituentType();
   makeBinary();
   assignCategories();

CCGBank Uses

Statistical Categorial Grammar Parsing

Julia’s own work

Identifying semantic roles

Various wide-coverage CCG parsers