Treebank-based Acquisition of LFG Resources for Chinese

Yuqing Guo
NCLT, DCU
26 Jul. 2006
Outline

- **Introduction**
- **Chinese Grammar**
  - Characteristics of Chinese Grammar
  - Chinese LFG
- **F-Structure Annotation Algorithm**
  - Experiments
  - Initial Analysis of results
- **Ongoing and future work**
Introduction

• Treebank-based Automatic Induction of Deep Grammar Resources for Chinese, Japanese, Arabic, Spanish, French, German and English
  – Development of large-scale deep unification grammars by hand is time-consuming and expensive
  – F-structure annotated tree-banks can be used for a parser to produce deep grammars
Chinese Language

- Written system
  - Chinese characters
    - pictogram:
      - pictogram: 上、下
      - ideogram: 上、下
      - logical aggregates: 明，众,
      - pictophonetic compounds: 妈
  - No space between words
    - difficult for segmentation
      - 球 拍 卖 完 了 vs. 球 拍 卖 完 了
        - racket sell out vs. ball auction finish
Chinese Grammar

• All words have only one grammatical form
  – no NUM, PER, TENSE inflection
    我正在看书 vs. 他看过这本书了
    I now read book he read have the CLS book LE
    ‘I’m reading the book now.’ ‘he has read the book.’
  – no agreement between subject and verb
  – Part-Of-Speech tagging
    他投资房地产 vs. 他 对 房地产 作 投资
    He invest real estate he in real estate make investment
    ‘He invested in real estate.’ ‘He made an investment in real estate.’
Chinese Grammar (cont.)

• A strong pro-drop tendency

– Subject

据说 明天 要 下雨
say tomorrow will rain.
‘It’s said that it’s going to rain tomorrow.’

– Predicate

• ADJP predicate

这 很 重要
It very important
‘It is very important.’

• NP or QP predicate

我家 三 个 孩子
My family three CLS child
‘my family has three children.’
## Sentence Elements & GFs

<table>
<thead>
<tr>
<th>Attributive</th>
<th>Subject</th>
<th>Adverbial</th>
<th>Predicate</th>
<th>Object</th>
<th>Complement</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>SUBJ/ADJ</td>
<td>ADJ/OBL</td>
<td>PRED</td>
<td>OBJ/OBJ2 COMP/XCOMP</td>
<td>ADJ/OBL</td>
</tr>
<tr>
<td>我们</td>
<td>our</td>
<td>公司</td>
<td>去年</td>
<td>实现 make</td>
<td>利润 profit</td>
</tr>
</tbody>
</table>

- Word order and categorial information play the main roles
1. 墙 上 挂 着 地图
   wall on hang ASP map
   ‘There is a map hanging on the wall.’

2. 那里 今天 下雨 了
   there today rain LE
   ‘It’s raining there today.’
1a. 这本书我看过关
This CLS book I read ASP
‘This book, I have read.’

1b. \[
\begin{aligned}
\text{PRED} & \quad \text{‘看<SUBJ, OBJ>’} \\
\text{TOPIC} & \quad [\text{PRED} \ ‘这本书’] \\
\text{SUBJ} & \quad [\text{PRED} \ ‘我’] \\
\text{OBJ} & \quad
\end{aligned}
\]

2a. 游泳我最拿手
swimming I the best
‘I’m the best at swimming.’

2b. \[
\begin{aligned}
\text{PRED} & \quad \text{‘拿手<SUBJ, OBJ>’} \\
\text{TOPIC} & \quad [\text{PRED} \ ‘游泳’] \\
\text{SUBJ} & \quad [\text{PRED} \ ‘我’] \\
\text{OBJ} & \quad
\end{aligned}
\]
3. 设备从安装到投产用了三个月

It took three months to install and run the equipment.
F-Structure Annotation

Coordination?

Y
Identify each coordinate

N
Head rules

Left-Right context rules

Catch-all and Clean up

LDD resolution

- Configurational information
- Categorial information
- Treebank functional tags
- Treebank co-indexing
### F-Structure Annotation Algorithm

- **Coordination**
  - Conjunction or coordinating punctuation
  - All children are in the same category

- **Head Rules**

<table>
<thead>
<tr>
<th>LHS</th>
<th>Direction</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJP</td>
<td>Right</td>
<td>JJ, ADJP</td>
</tr>
<tr>
<td>ADVP</td>
<td>Right</td>
<td>AD, CS, ADVP</td>
</tr>
<tr>
<td>CLP</td>
<td>Right</td>
<td>M, CLP</td>
</tr>
<tr>
<td>...</td>
<td>Right</td>
<td>...</td>
</tr>
<tr>
<td>IP</td>
<td>Left</td>
<td>VP, IP</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
F-Structure Annotation Algorithm

- Left-Right Context Rules

<table>
<thead>
<tr>
<th>NP</th>
<th>Left context</th>
<th>Head</th>
<th>Right context</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJP, ADVP, CP, DNP, IP...</td>
<td>↓ ∈ ↑ ADJ</td>
<td>NN, NR, NT, PN, NP, QP: ↑ = ↓</td>
<td>QP: ↓ ∈ ↑ ADJ</td>
</tr>
<tr>
<td>CLP: ↑ SPEC: QUANT = ↓</td>
<td></td>
<td></td>
<td>ETC: ↓ ∈ ↑ COORD</td>
</tr>
</tbody>
</table>

- Catch-all & Clean up
  - Functional tags
  - Default rules for remaining nodes
  - Correct the over-generalisation
Long-Distance Dependency

```
WHNP-1
[↑TOPIC_REL=↓]

-NONE- *OP*

CP
[↓∈↑ADJ]

NP
[↑=↓]

-NONE- *T-1

IP
[↑=↓]

NP
[↑SUBJ=↓]

ADVP
[↓∈↑ADJ]

already

PRED
'this business'

PRED
'开业(SUBJ)'

SUBJ
PRED

ADJ

de

VP
[↑=↓]

DE
[↑de=’+’]

NP
[↑=↓]

NN
[↑=↓]

DEC
[↑de=’+’]

foreign business

PRED
'this business'

ADJ

{TOPIC_REL
PRED
SUBJ
ADJ
de

PRED
'已'

}
# Experiment Results - Quantitative Evaluation

<table>
<thead>
<tr>
<th>CTB2</th>
<th>Sentences</th>
<th>Percent(%)</th>
<th>CTB5</th>
<th>Sentences</th>
<th>Percent(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4133</td>
<td>98.8047</td>
<td>1</td>
<td>17699</td>
<td>94.2088</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>0.9563</td>
<td>2</td>
<td>849</td>
<td>4.5191</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.1195</td>
<td>3</td>
<td>123</td>
<td>0.6547</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.0956</td>
<td>4</td>
<td>43</td>
<td>0.2289</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.0239</td>
<td>5</td>
<td>20</td>
<td>0.1065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>12</td>
<td>0.0639</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>5</td>
<td>0.0266</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td>35</td>
<td>0.1863</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>1</td>
<td>0.0053</td>
</tr>
<tr>
<td>total</td>
<td>4183</td>
<td></td>
<td>total</td>
<td>18787</td>
<td></td>
</tr>
</tbody>
</table>
## Experiment Results - Qualitative Evaluation

<table>
<thead>
<tr>
<th>Pred-only GFs</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-score (%)</th>
<th>Other features</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjunct</td>
<td>90.79</td>
<td>85.19</td>
<td>87.90</td>
<td>de</td>
<td>95.00</td>
<td>58.46</td>
<td>72.38</td>
</tr>
<tr>
<td>comp</td>
<td>52.63</td>
<td>38.46</td>
<td>44.44</td>
<td>di</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>coord</td>
<td>46.55</td>
<td>43.55</td>
<td>45.00</td>
<td>noun_type</td>
<td>99.80</td>
<td>90.69</td>
<td>95.03</td>
</tr>
<tr>
<td>det</td>
<td>100.00</td>
<td>33.33</td>
<td>50.00</td>
<td>number_type</td>
<td>100.00</td>
<td>88.89</td>
<td>94.12</td>
</tr>
<tr>
<td>numeral</td>
<td>54.55</td>
<td>92.31</td>
<td>68.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obj</td>
<td>86.44</td>
<td>91.62</td>
<td>88.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obj2</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>obl</td>
<td>66.67</td>
<td>66.67</td>
<td>66.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quant</td>
<td>86.36</td>
<td>57.58</td>
<td>69.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spec</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>subj</td>
<td>50.37</td>
<td>53.13</td>
<td>51.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>topic_rel</td>
<td>94.12</td>
<td>82.05</td>
<td>87.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xcomp</td>
<td>66.67</td>
<td>62.50</td>
<td>64.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pred-only</td>
<td>75.58</td>
<td>72.68</td>
<td>74.10</td>
<td>Overall</td>
<td>83.87</td>
<td>78.05</td>
<td>80.85</td>
</tr>
</tbody>
</table>
Error Analysis

- Treebank bracketing error
  - A level of phrasal category is missing:
    e.g. IP -> IP IP vs. IP -> NP VP IP
  - Functional tag error

- Flat structure
  e.g. NP -> NN NN NN NN

- Other Ambiguities
  e.g. IP -> NP NP VP
Future works

- A larger gold-standard f-structure evaluation set
- From dependency tree to f-structure
- Propbank with word semantic forms
Thanks!
&
Any Questions?