PS to DS Conversion of the Chinese Treebank

Nianwen Xue
CLEAR and Linguistics
University of Colorado

Dec 7, 2007
TLT’07, Bergen, Norway
Chinese Treebank (1998 - ?)
(Xue, Xia, Chiou, Palmer 2005, JNLE)

邱福栋 Fu-Dong Chiou
蒋自新 Zixin Jiang
石美莎 Martha S. Palmer
夏 飞 Fei Xia
薛念文 Nianwen Xue
张美玉 Meiyu Chang
CTB: overview

http://verbs.colorado.edu/chinese

• Started at Penn, being continued at University of Colorado at Boulder
• Supported by DOD, NSF, DARPA
• Now a nearly 1M word Chinese corpus
  – Xinhua newswire, Sinorama news magazine, HK news, broadcast news, broadcast conversation
  – Segmented, POS-tagged, syntactically bracketed
  – Phrase structure annotation
• Additional layers of annotation
  – Propbank/Nombank, Discourse relations
• Desired annotation
  – Timebank, more discourse relations
<table>
<thead>
<tr>
<th>Version</th>
<th>Year</th>
<th>Quantity (words)</th>
<th>Source</th>
<th>Propbank/Nombank</th>
<th>Discourse annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTB1.0</td>
<td>2001</td>
<td>100K</td>
<td>Xinhua</td>
<td>yes</td>
<td>Pilot</td>
</tr>
<tr>
<td>CTB3.0</td>
<td>2003</td>
<td>250K</td>
<td>+HK News</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>CTB4.0</td>
<td>2004</td>
<td>400K</td>
<td>+Sinorama</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>CTB5.0</td>
<td>2005</td>
<td>500K</td>
<td>+Sinorama</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>CTB6.0</td>
<td>2007</td>
<td>780K</td>
<td>+BN</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>CTB7.0</td>
<td>2008</td>
<td>950K</td>
<td>+BC</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>CTB8.0??</td>
<td>2009??</td>
<td>1.1M??</td>
<td>+??</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
Uses

- **Segmentation**
- **POS tagging**
  - Tseng et al 2005, Hillard et al 2006, Xia et al 2006, ...
- **BaseNP chunking**
  - Liang et al 2006, Xu et al 2006, Chen et al 2006...
- **Empty category recovery**
  - Zhao and Ng 2007,
- **Grammar extraction**
  - Xia 2000, Burke et al 2004, Guo et al 2007,
- **Classifier Assignment**
  - Guo and Zhong 2005
- **Machine Translation**
  - Wang, Collins and Koehn 2007,
More uses

• Constituent structure parsing

• Dependency structure parsing
A Penn Treebank Example: no arg/adj distinction

Chinese Treebank modification

predication  adjunction  complementation
Another Penn Treebank Example

The company

expects

VP1

VP2

to

VP3

obtain regulatory approval

and

complete transaction

by year-end
Chinese Treebank modification

coordination  adjunction
Principles are hard to resist

Coordination  Adjunction
One grammatical relation per pair of brackets

Three primitive grammatical relations

Grammatical relation: head/dependent
Complementation (left-headed)
Implicit information

XP
  \- X-head
  \- YP
  \- \{ZP\}

DP
  \- DeTerminer
  \- QP

VP
  \- VV
  \- NP

PP
  \- P
  \- NP
Complementation (right-headed)

- XP
  - YP
  - X

- CP
  - IP
  - DEC

- DNP
  - NP
  - DEG

- LCP
  - NP/IP
  - LoCalizer
Implicit information

XP
  YP
  X-head

CP
  IP
  DEC

DNP
  NP
  DEG

LCP
  NP/IP
  LoCalizer
Adjunction

XP

{YP}  XP  {ZP}

VP

ADVP  VP  ADJP

ADVP  ADJP
Implicit information

XP

\{YP\}   \textbf{XP-Head}   \{ZP\}

VP

ADVP    VP

ADJP

ADVP    ADJP
Coordination

XP
  /   \
/     \ 
XP     CC
   /   \
VP     VP
  /     \
VP     CU
 /       \
VP     VP
 /         \
NP     CC
   /       \
NP
Implicit information

XP
/   \\  
XP  CC-Head??  XP
/   \\      \\/   \  
VP    VP         NP
    /   \\     /   \\  
VP    CC   VP    NP
    /     \    /       \\/  
VP    VP   PU  VP
    /   \\   / \\  
NP    NP  CC  NP
Why is coordination especially difficult for dependency tree representation

- **Properties of dependency tree**
  - Explicit commitment to a lexical head under the assumption that there is one head for each dependent
  - Doesn’t have a non-terminal to mediate the external distributional properties

- **Properties of coordinate structure**
  - Thematic head that determines external distributional properties doesn’t converge with functor that holds the conjuncts together

- **Note**
  - Lack of arg/adj distinction does not matter for dependency structure because it doesn’t differentiate adj/arg by convention (all modifiers).
She was untrained and in one botched job killed a client.
Conjunction as head

- Making coordination the head fails to capture the external distributional properties of the coordinate construction
+ Possible solution: make this dependency transparent: dependencies with the coordinate conjunction inherited by its children (c.f. Prague Dependency Treebank)
She was untrained and in one botched job killed a client.
She was untrained and in one botched job killed a client.
First conjunct as head

+ Addresses the external dependencies of the coordinate construction

- But still problematic when it is the head that takes some dependent
Head table approach to PS-DS conversion

- Mark the head child of each node in a phrase structure, using the head table, which consists of list of head rules
- For each node, make the lexical head of each non-head child depend on the lexical head of the head child

<table>
<thead>
<tr>
<th></th>
<th>VP: VBD, VP, ...</th>
<th>NP: NN, ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP: P, ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>....</td>
</tr>
</tbody>
</table>
She was untrained and in one botched job killed a client.

Head rule: VP ~ l VP; l VBD;l VBN;l MD;l VBZ;l VB; l ADJP; l NN; l NNS; l NP
How it works (complementation)

She was untrained and in one botched job killed a client

Head rule: VP ~ 1 VP;1 VBD;1 VBN;1 MD;1 VBZ;1 VB; 1 ADJP; 1 NN; 1 NNS; 1 NP
How it works (coordination)

She was untrained and in one botched job killed a client

Head rule: VP ~ l, VP;l VBD;l VBN;l MD;l VBZ;l VB; l ADJP; l NN; l NNS; l NP
She was untrained and in one botched job killed a client.

Head rule: VP ~ 1 VP; 1 VBD; 1 VBN; 1 MD; 1 VBZ; 1 VB; 1 ADJP; 1 NN; 1 NNS; 1 NP
An alternative approach

- **Algorithm:**
  - Identify the linguistic structure for each constituent (complementation, adjunction or coordination)
  - Identify the head child of each constituent in a phrase structure
  - For each constituent, make the lexical head of each non-head child depend on the lexical head of the head child
- **Exploiting the richer structural information**
  - Requires clear distinctions between different structures
  - Three primitive linguistic structures in the CTB
How big a deal is this?

- Calculate the agreement between the head table and structural approach
- The structural approach assumes that the coordinating conjunction is the head (taking the lesser of two evils)
- Used Penn2Malt to do the conversion for the head table approach ([http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html](http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html))
- Experiments done on 250K words of the Chinese Treebank
- Agreement = overlapping edges / total edges
- Unlabeled agreement
- The agreement between the two approaches is…
- 60%
Is the head table extensible?

- Possible if the head is a coordinating conjunction, just a bit clumsy
- Problematic if the head is a punctuation mark, etc..
Summary

• An overview of the Chinese Treebank representation scheme
• Coordinate structures account for a high proportion of the dependency relations (and therefore it’s important to get it right)
• There is enough structure in the CTB to allow more flexibility in determining the head than the head table approach can fully exploit