Treebank-based Acquisition of LFG Resources for Chinese

Yuqing Guo
NCLT, DCU
26 Jul. 2006

Outline

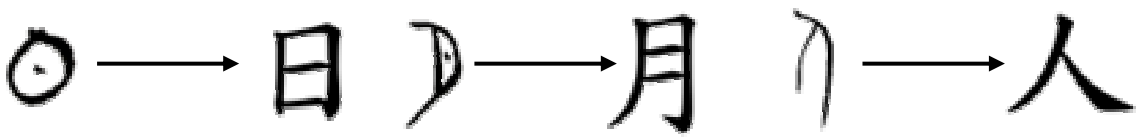
- **Introduction**
- **Chinese Grammar**
  - Characteristics of Chinese Grammar
  - Chinese LFG
- **F-Structure Annotation Algorithm**
  - Experiments
  - Initial Analysis of results
- **Ongoing and future work**

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:



    - 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 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

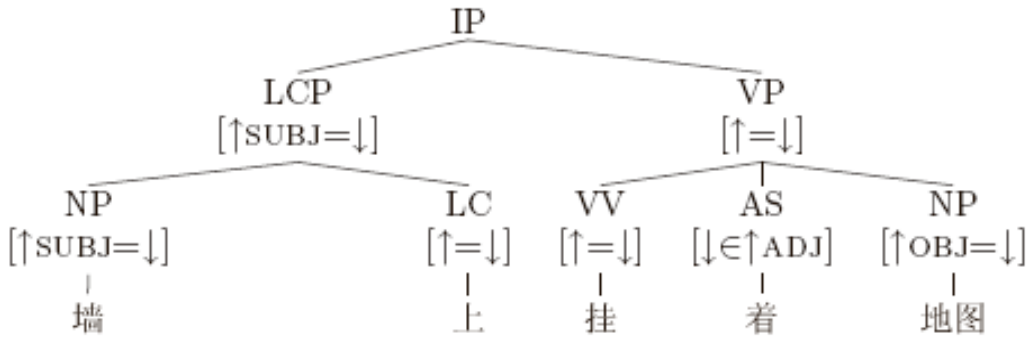
Attributive	Subject	Adverbial	Predicate	Object	Complement
ADJ	SUBJ/ ADJ	ADJ/OBL	PRED	OBJ/OBJ2 COMP/ XCOMP	ADJ/OBL
我们 our	公司 company	去年 last year	实现 make	利润 profit	五百万美元 5 million dollars

- Word order and categorial information play the main roles

Locative & Temporary Subject

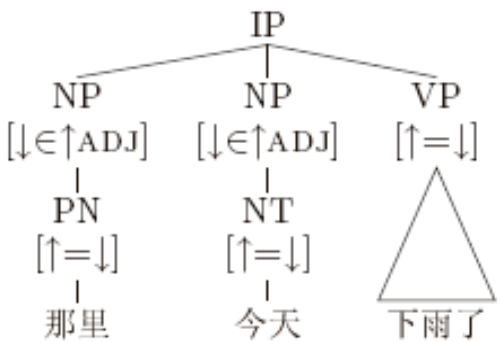
1. 墙 上 挂 着 地图
   wall on hang ASP map
   ‘There is a map hanging on the wall.’

$$\begin{bmatrix} \text{PRED} & \text{'挂⟨SUBJ, OBJ⟩'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'上'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'墙'} \end{bmatrix} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'地图'} \end{bmatrix} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'着'} \end{bmatrix} \right\} \end{bmatrix}$$



2. 那里 今天 下雨 了
   there today rain LE
   ‘It’s raining there today.’

$$\begin{bmatrix} \text{PRED} & \text{'下雨⟨SUBJ⟩'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \end{bmatrix} \\ \text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'那里'} \end{bmatrix} \begin{bmatrix} \text{PRED} & \text{'今天'} \end{bmatrix} \right\} \end{bmatrix}$$



Topic

1a. 这 本 书 我 看 过
This CLS book I read ASP
'This book, I have read.'

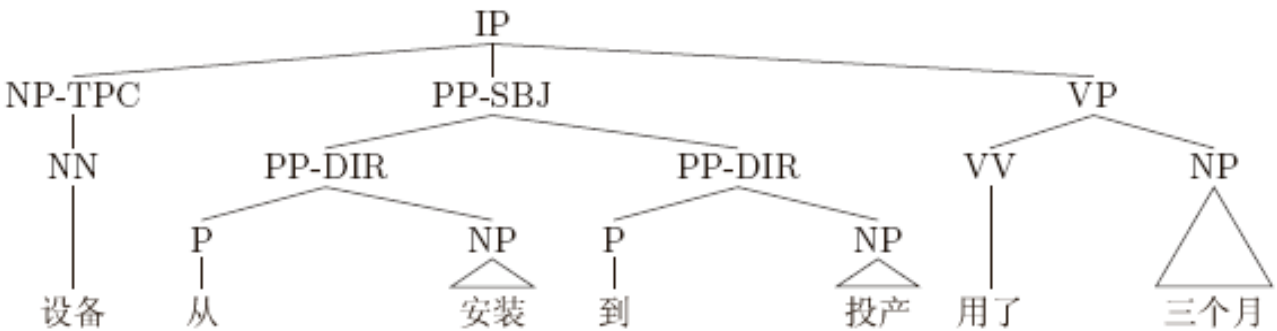
1b.
PRED '看<SUBJ, OBJ>'
TOPIC [PRED '这本书']
SUBJ [PRED '我']
OBJ

2a. 游泳 我 最拿手
swimming I the best
'I'm the best at swimming.'

2b.
PRED '拿手<SUBJ, OBJ>'
TOPIC [PRED '游泳']
SUBJ [PRED '我']
OBJ

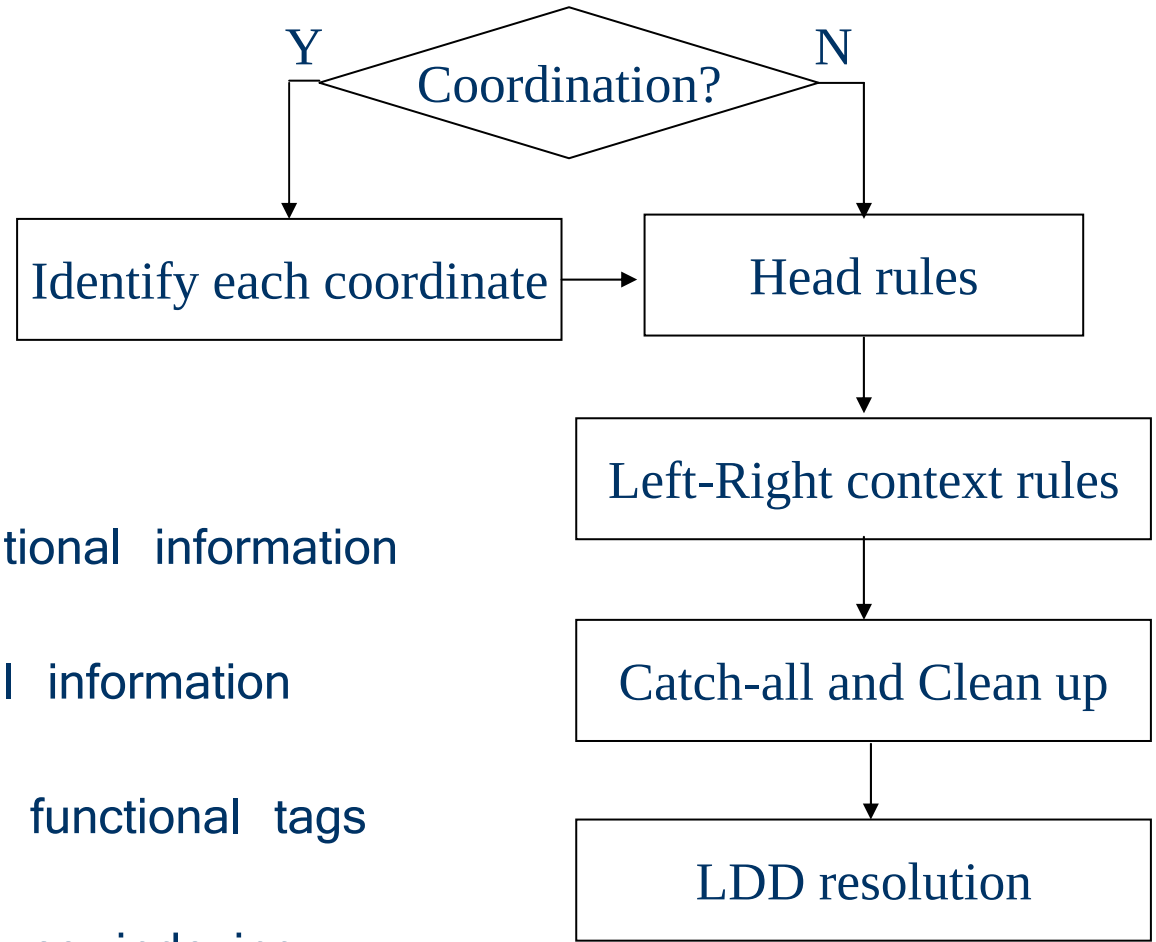
Topic (cont.)

3. 设备 从 安装 到 投产 用 了 三 个 月
equipment from install to run take LE three CLS month
'It took three months to install and run the equipment.'



$$\begin{bmatrix} \text{PRED} & \text{'用⟨SUBJ, OBJ⟩'} \\ \text{TOPIC} & \begin{bmatrix} \text{PRED} & \text{'设备'} \end{bmatrix} \\ \text{SUBJ} & \begin{bmatrix} \text{COORD} & \left\{ \begin{bmatrix} \text{PRED} & \text{'从⟨OBJ⟩'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'安装'} \end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{PRED} & \text{'到⟨OBJ⟩'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'生产'} \end{bmatrix} \end{bmatrix} \right\} \\ \text{COORD\_FORM} & \text{'null'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'三个月'} \end{bmatrix} \end{bmatrix}$$

F-Structure Annotation



- 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
  - Levy & Manning (2003) head-lexicalisation rules

LHS	Direction	RHS
ADJP	Right	JJ, ADJP
ADVP	Right	AD, CS, ADVP
CLP	Right	M, CLP
…	…	…
IP	Left	VP, IP
…	…	…

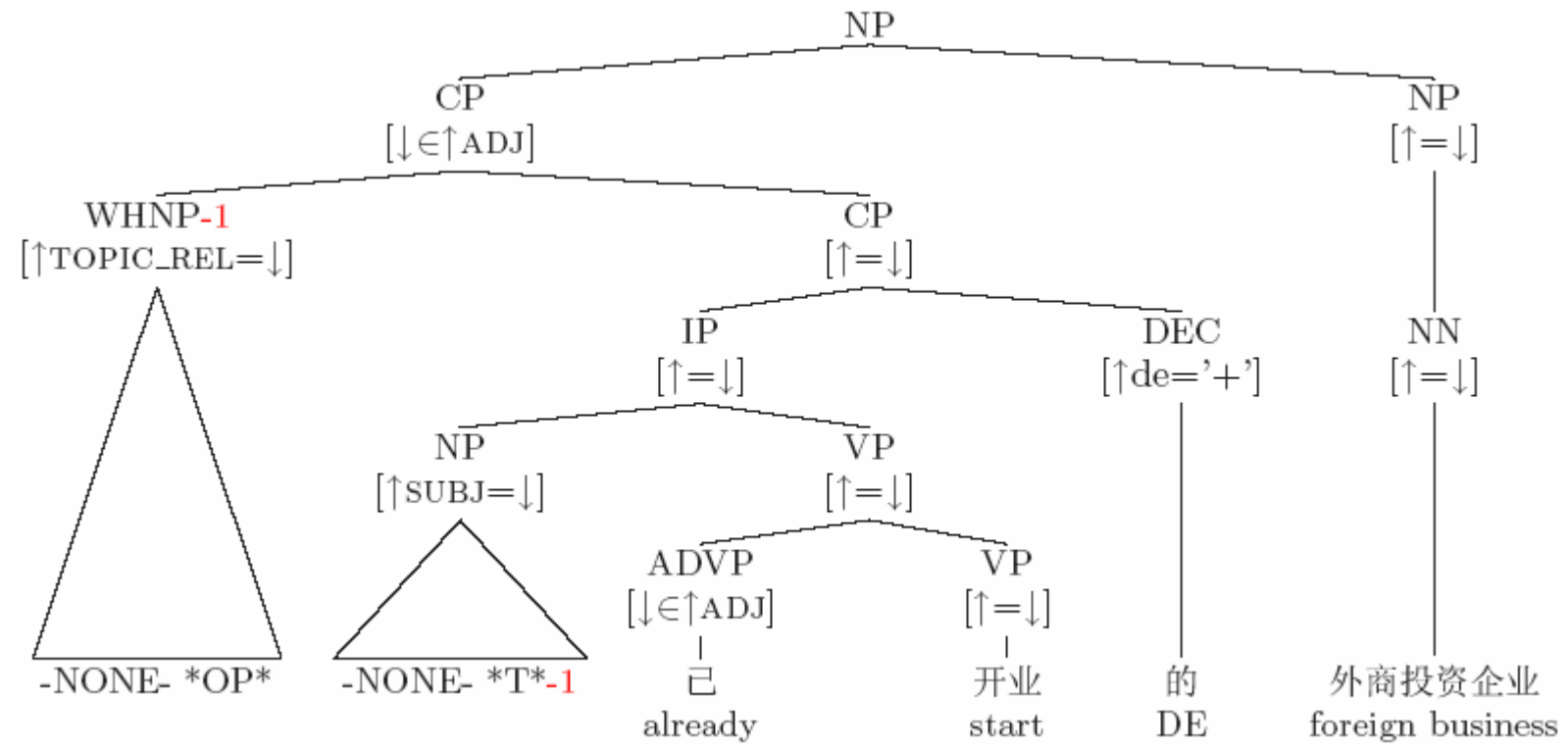
F-Structure Annotation Algorithm

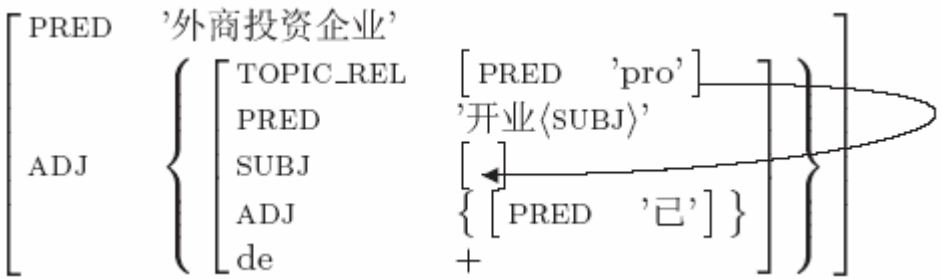
- Left-Right Context Rules

NP	Left context	Head	Right context
	ADJP, ADVP,CP, DNP,IP…: ↓ ∈ ↑ ADJ CLP: ↑ SPEC:QUANT= ↓	NN, NR, NT, PN, NP, QP: ↑ = ↓	QP: ↓ ∈ ↑ ADJ ETC: ↓ ∈ ↑ COORD

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

Long-Distance Dependency





Experiment Results - Quantitative Evaluation

CTB2	Sentences	Percent(%)	CTB5	Sentences	Percent(%)
1	4133	98.8047	1	17699	94.2088
2	40	0.9563	2	849	4.5191
3	5	0.1195	3	123	0.6547
4	4	0.0956	4	43	0.2289
5	1	0.0239	5	20	0.1065
			6	12	0.0639
			7	5	0.0266
			9	35	0.1863
			10	1	0.0053
total	4183		total	18787	

Experiment Results - Qualitative Evaluation

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

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?