

Recovering Non-Local Dependencies for Chinese

Yuqing Guo¹, Haifeng Wang², Josef van Genabith^{1,3}

¹NCLT, School of Computing, Dublin City University, ³IBM Center for Advanced Studies, Dublin, Ireland

²Toshiba (China) Research and Development Center, Beijing, China

Introduction

Non-Local Dependencies (NLDs)

Linguistic phenomena permit a constituent in one position (*antecedent*) to bear the grammatical function associated with another position (*trace*).

Motivation

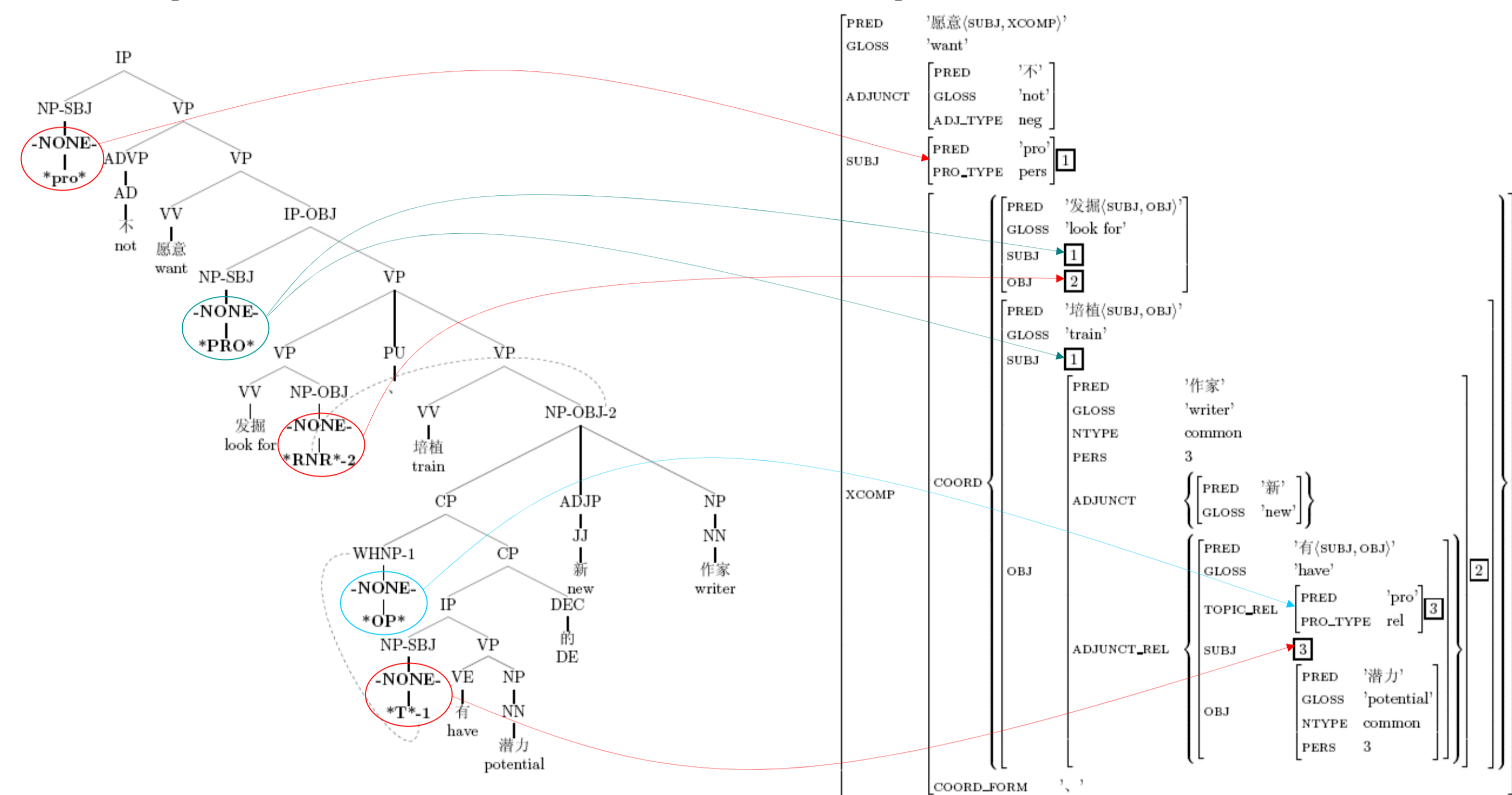
- Recovering NLDs is crucial to the accurate and complete determination of predicate-argument structures or deep dependencies
- Most of the state-of-the-art broad coverage statistical parsers do NOT capture NLDs
- To date, the research has focused almost entirely on English

NLDs in Chinese

- Null Elements: Empty relative pronouns
- Locally Mediated Dependencies: Short-Bei construction, Raising & Control constructions
- Long-Distance Dependencies: Wh-traces, Topicalisation, Coordination, Pro-drop situations etc.

Representation of NLDs in CTB & LFG

不愿意 发掘、培植有潜力的新作家
not want look for and train have potential DE new writer
'(People) don't want to look for and train new writers who have potential.'

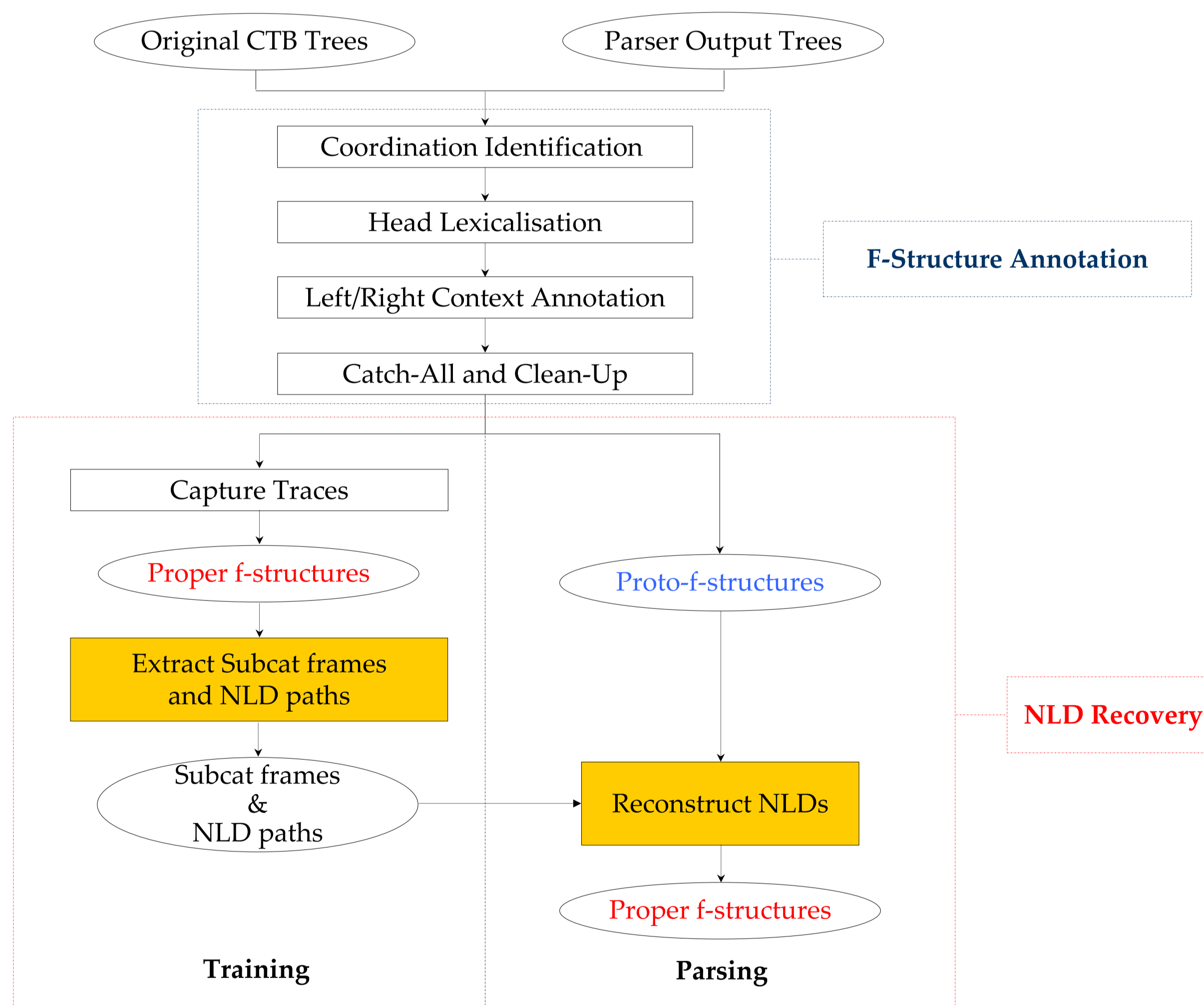


NLDs are represented as empty categories/traces and (for some of them) coindexation with antecedents in Penn Chinese Treebank (CTB)

NLDs are represented as reentrancies in LFG f-structures

NLD Recovery in LFG

System Architecture



NLD Resources

✦ Probability of NLD paths p linking traces and antecedents conditioned on trace t :

$$P(p|t) = \frac{\text{count}(p,t)}{\sum_{i=1}^n \text{count}(p_i,t)}$$

✦ Probability of subcat frames s conditioned on word w and its syntactic features:

$$P(s|w, w_feats) = \frac{\text{count}(s, w, w_feats)}{\sum_{i=1}^n \text{count}(s_i, w, w_feats)}$$

Trace (Path)	Probability	Word:POS-GF(Subcat Frames)	Probability
adjunct(out-adjunct:in-topic_rel)	0.9018	有:VE-adj_rel([subj, obj])	0.6769
adjunct(out-adjunct:out-coord:in-topic_rel)	0.0192	有:VE-adj_rel([subj, comp])	0.1531
adjunct(NULL)	0.0128	有:VE-adj_rel([subj])	0.0556
...
obj(out-obj:in-topic_rel)	0.7915	有:VE-comp([subj, obj])	0.4805
obj(out-obj:out-coord:in-coord:in-obj)	0.1108	有:VE-comp([subj, comp])	0.2587
...
subj(NULL)	0.3903	有:VE-top([subj, comp])	0.4397
subj(out-subj:in-topic_rel)	0.2092	有:VE-top([subj, obj])	0.3510
...

Methods

The Core Algorithm

- Traverse f-structure f inside-out
- Insert a dislocated argument t at sub-f-structure h whose local PRED is w , iff
 - t is not present at h yet
 - h together with t is locally complete and coherent with regard to w 's subcat frames s
- Traverse f starting from t along the NLD path p and link t to its antecedent a if p ends with a ; or leave t as PRO if p is a NULL path
- Rank all resolutions by: $P(s|w, w_feats) \times \prod_{j=1}^m P(p|t_j)$

The Hybrid Fine-Grained Strategy

- Applying a few simple heuristic rules to insert empty PRED for verbal coordination constructions and to insert empty relative pronouns for relativisation
- Inserting an empty node labelled as SUBJ for Locally Mediated Dependencies and relating it to the upper-level SUBJ or OBJ accordingly
- Resolving wh-traces in relative clauses including modifier antecedents of TOPIC & ADJUNCT, by using subcat frames and NLD paths conditioned on antecedents ((Cahill et al., 2004)'s algorithm)
- Using our modified algorithm to recover the remaining (and majority) NLD types

Experiments and Evaluation

LFG F-structure Annotation

Dependencies	Precision	Recall	F-Score
gold standard CTB trees	95.60	95.82	95.71
Bikel's parser output trees	74.37	73.15	73.75

F-structure-based NLD Recovery

The Basic Model

CTB trees	Trace Insertion			Antecedent Recovery		
	Precision	Recall	F-Score	Precision	Recall	F-Score
(Cahill et al., 2004)	95.98	57.86	72.20	90.16	54.35	67.82
Our Basic Model	92.44	91.28	91.85	63.12	62.33	62.72
Subject Path Constraint	92.16	91.36	91.76	75.96	75.30	75.63

- Cahill et al. 2004: The algorithm resolving wh-traces for relativisation
- Basic Model: Our algorithm recovering all traces associated with governable grammatical functions
- Subject Path Constraint: Our basic algorithm with subject path constraint to weight non-empty paths for SUBJ

The Hybrid Fine-Grained Model

CTB trees	Basic Model			Fine-Grained Model		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Trace Insertion						
Overall	92.16	91.36	91.76	92.86	91.45	92.15
SUBJ	92.95	97.81	95.32	94.38	97.81	96.06
OBJ	65.28	64.98	65.13	78.95	55.30	65.04
ADJUNCT	0.0	0.0	0.0	38.24	25.49	30.59
TOPIC	0.0	0.0	0.0	33.33	35.14	34.21
TOPIC_REL	99.85	99.39	99.62	99.85	99.39	99.62
COORD	90.00	100.00	94.74	90.00	100.00	94.74
Antecedent Recovery						
Overall	75.96	75.30	75.63	84.92	83.64	84.28
SUBJ	66.93	70.42	68.63	81.61	84.57	83.06
OBJ	61.57	61.29	61.43	75.66	53.00	62.33
ADJUNCT	0.0	0.0	0.0	38.24	25.49	30.59
TOPIC	0.0	0.0	0.0	33.33	35.14	34.21
TOPIC_REL	99.85	99.39	99.62	99.85	99.39	99.62
COORD	90.00	100.00	94.74	90.00	100.00	94.74

Breakdown by major grammatical functions for evaluation of trace insertion and antecedent recovery on CTB trees stripped of empty nodes and coindexation

Better Training for Parser output

- Reparse the training data by Bikel's parser using 10-fold cross-validation
- Convert reparsed data and original data into f-structures
- Restore traces and coindexation on reparsed data by comparing with original data

Bikel's Parser Output Tree	Trace Insertion			Antecedent Recovery		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Hybrid Modal	64.07	62.37	63.21	54.53	53.08	53.79
Hybrid+better-train	67.29	62.33	64.71	56.88	52.69	54.71

Conclusion

- ✦ An algorithm for recovering Non-Local Dependencies in Chinese on CTB-based automatically annotated LFG f-structures, using
 - probabilities of NLD paths conditioned on grammatical function associated with the trace
 - probabilities of subcat frames conditioned on combination of the word form and its syntactic features
- ✦ A hybrid strategy to recover different NLD types in light of their diverse linguistic properties
- ✦ A better training method to improve the similarity of training material to parser output trees