

Trebank-Based Acquisition of LFG Resources for Chinese

Yuqing Guo¹, Josef van Genabith^{1,2}, and Haifeng Wang³

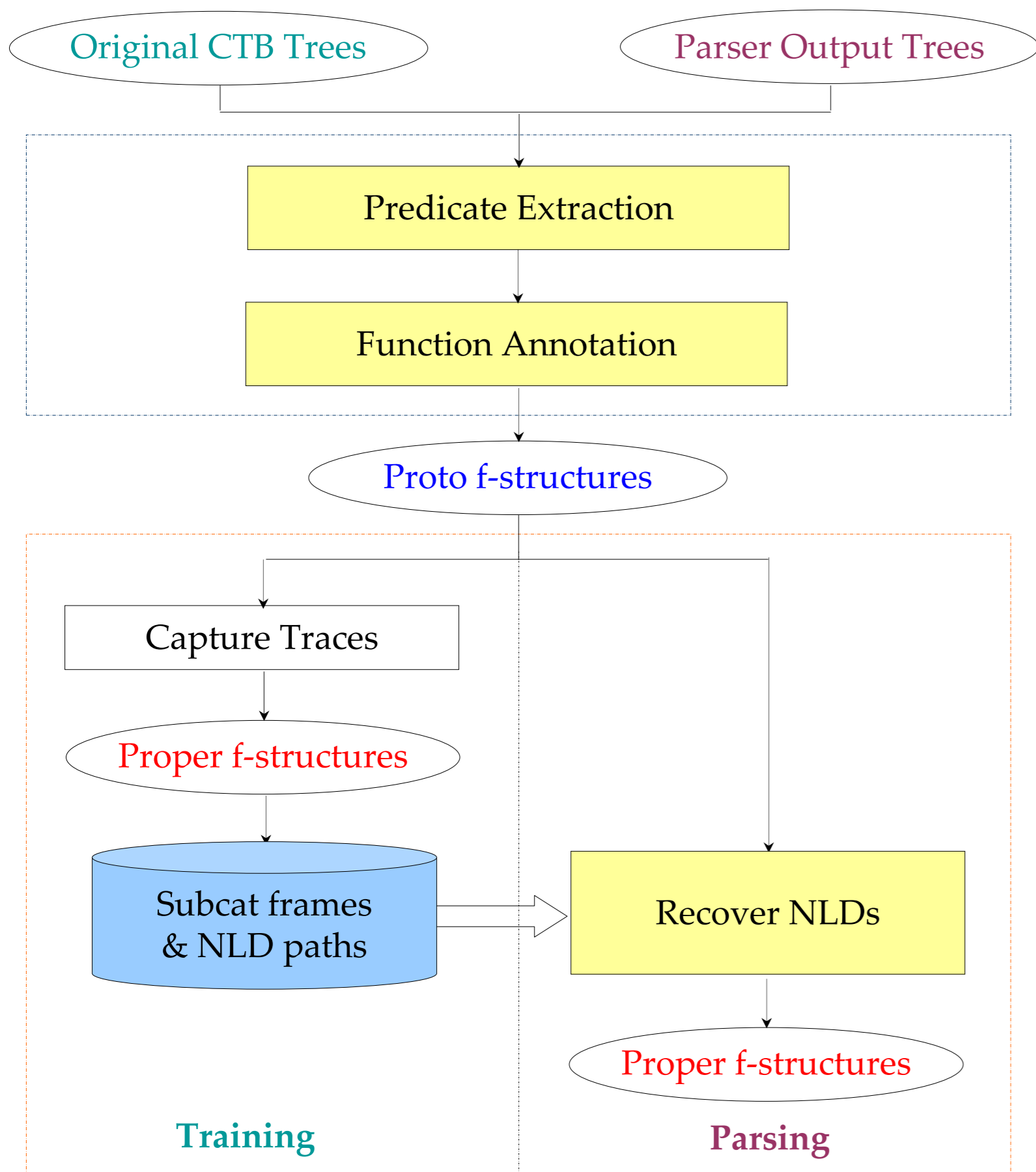
¹NCLT, School of Computing, Dublin City University

²IBM Center for Advanced Studies, Dublin, Ireland

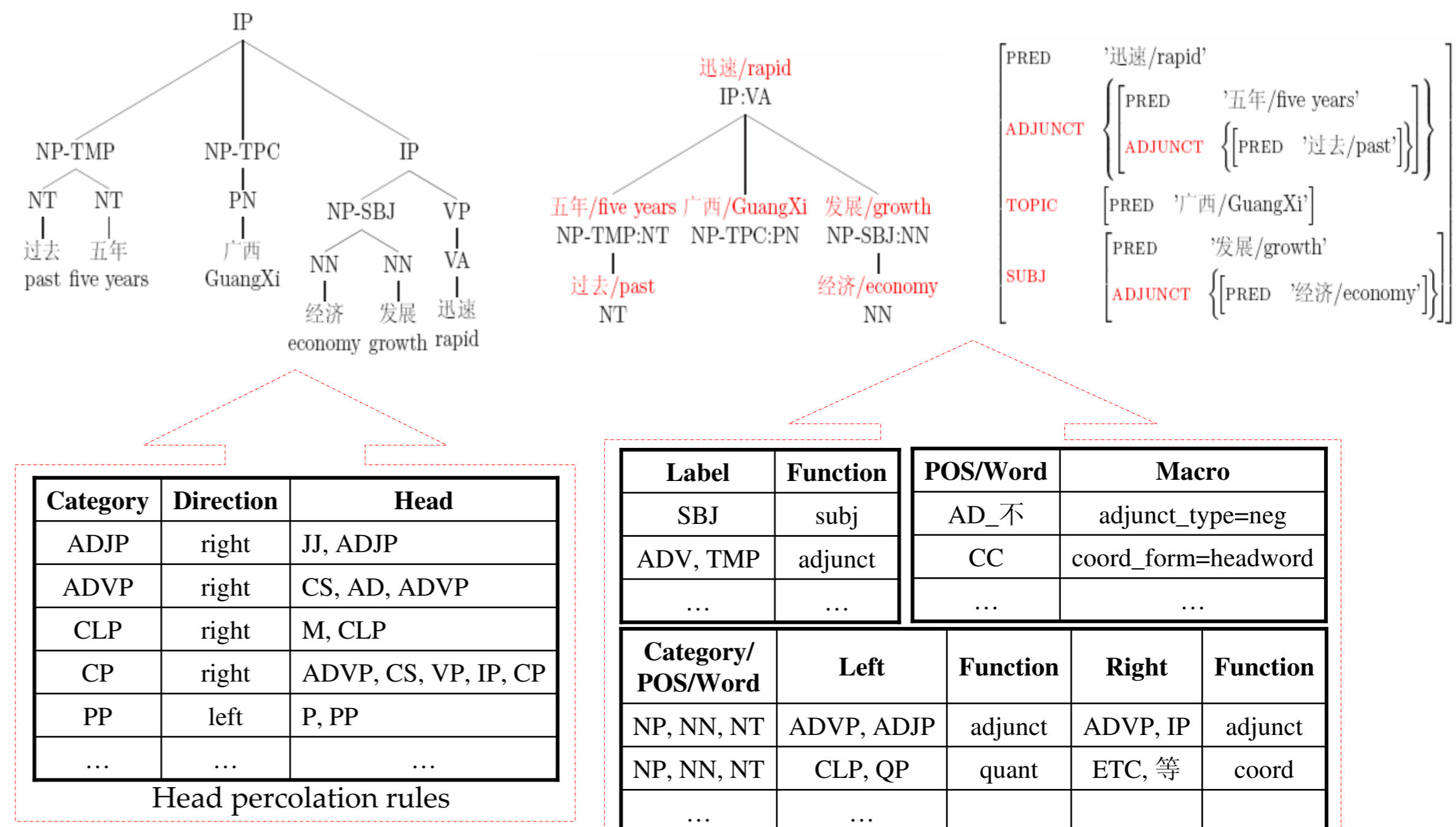
³Toshiba (China) Research and Development Center, Beijing, China



System Architecture



F-Structure Annotation



Results

	CTB Trees			Bikel Parser Output		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Preds Only	93.68	94.93	94.30	73.55	65.05	69.04
All GFs	95.25	96.75	96.00	84.00	71.77	77.40

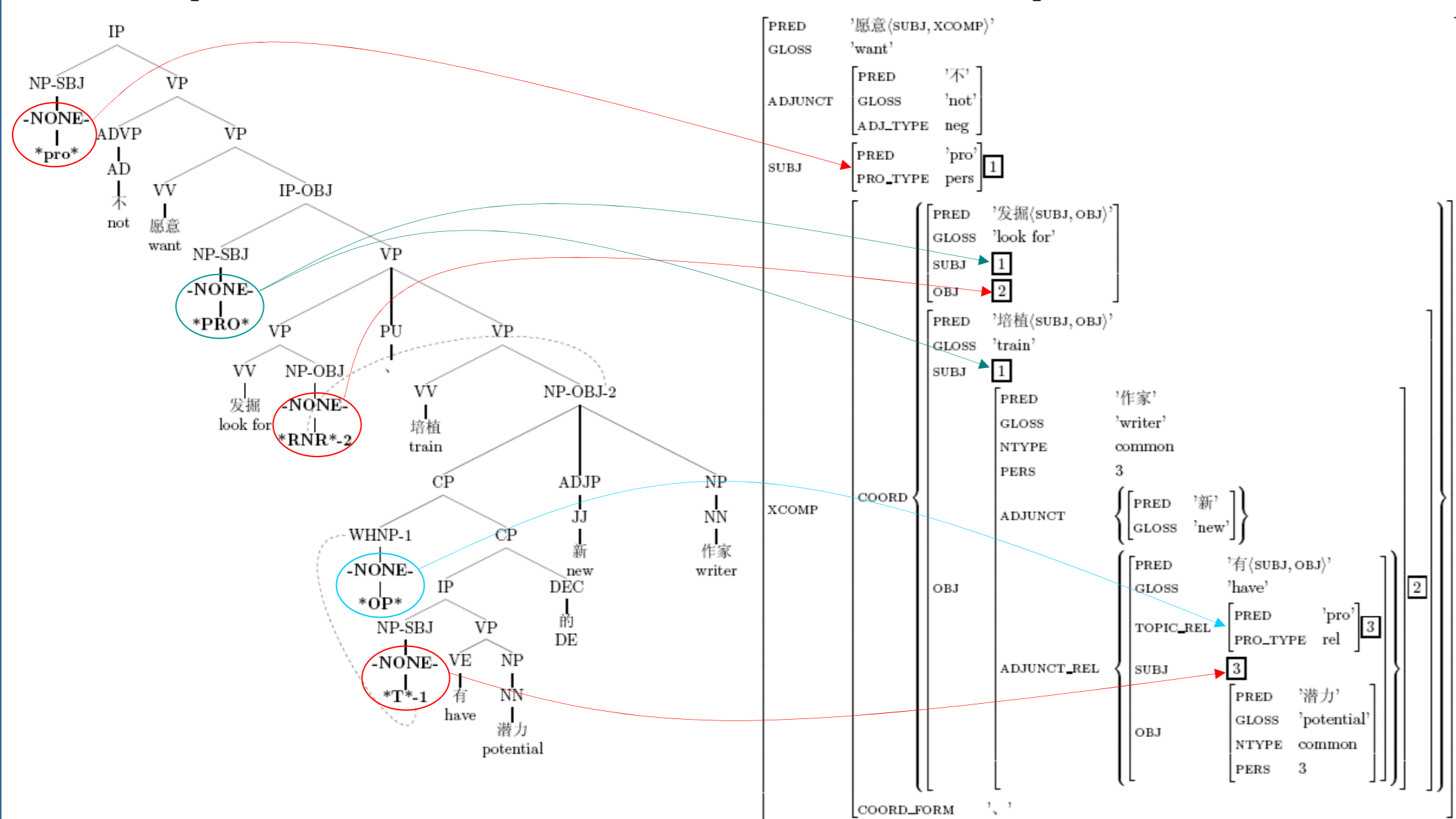
Evaluation of f-structure generation against 200-sentence gold standard

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.'



Non-Local Dependency Recovery

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)	Prob.	Word:POS-GF(subcat frames)	Prob.
adj(out-adj:in-topic_rel)	0.9018	有:VE-adj_rel([subj, obj])	0.6769
adj(out-adj:out-coord:in-topic_rel)	0.0192	有:VE-adj_rel([subj, comp])	0.1531
adj(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

Results

	Precision	Recall	F-Score
Insertion	92.16	91.36	91.76
Recovery	75.96	75.30	75.63

Evaluation of NLD recovery against 1,046 CTB trees stripped off coindexation

Result

+NLD res.	Precision	Recall	F-Score
Preds Only	71.91	70.81	71.36
All GFs	80.41	79.61	80.01

Evaluation of f-structure annotation with NLD recovery against 200-sentence gold standard

Conclusions

- A robust and accurate algorithm for automatic f-structure annotation of Penn Chinese Treebank 5.1
- Non-local dependencies are recovered on automatically generated f-structures
- LFG resources for Chinese are acquired from the proper f-structures