Improving Tree-to-Tree Translation with Packed Forests

Yang Liu, Yajuan Lü, and Qun Liu

Institute of Computing Technology,

Chinese Academy of Sciences

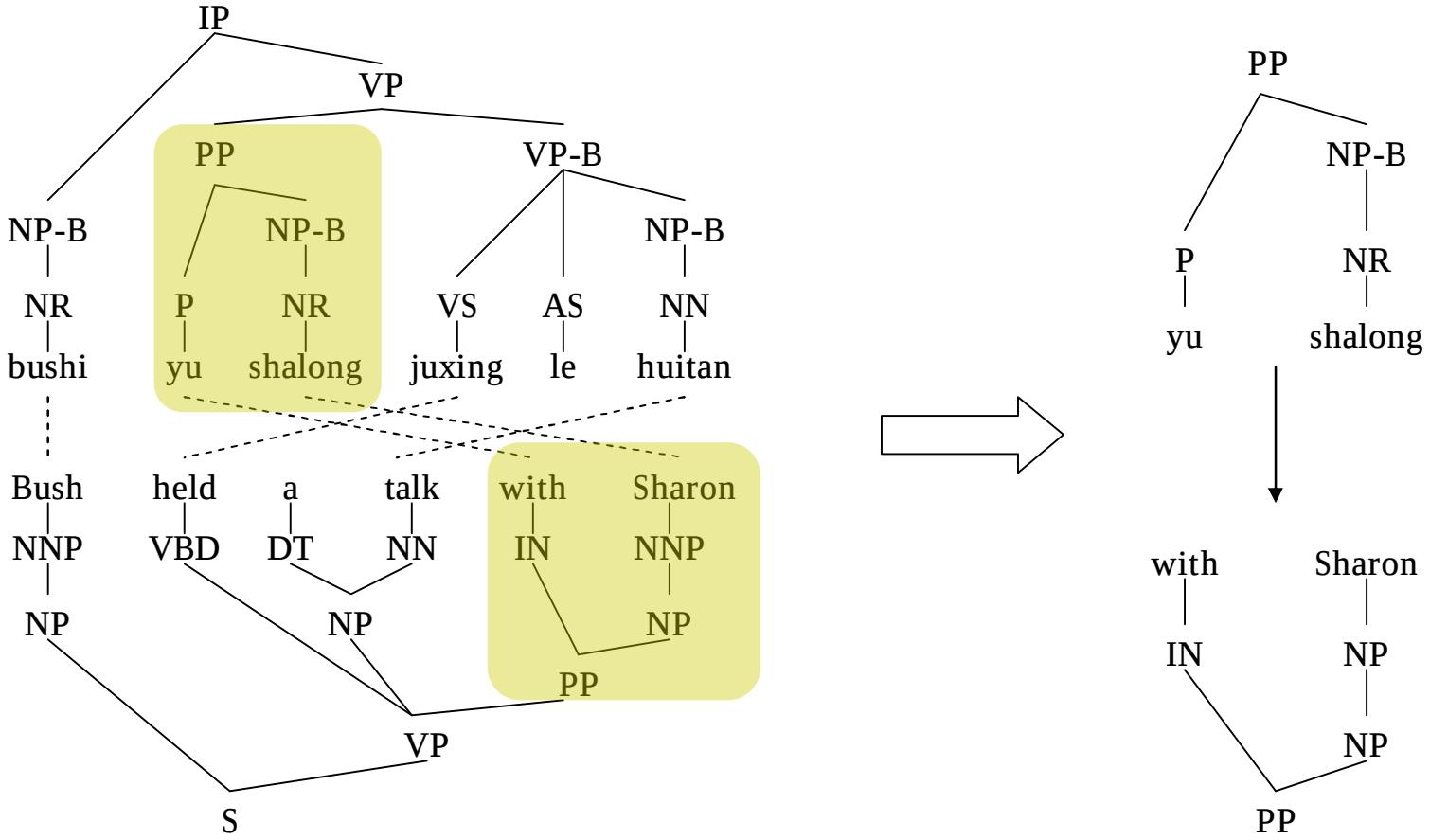
{yliu,lvyajuan,liuqun}@ict.ac.cn

中科院计算所

INSTITUTE OF COMPUTING TECHNOLOGY

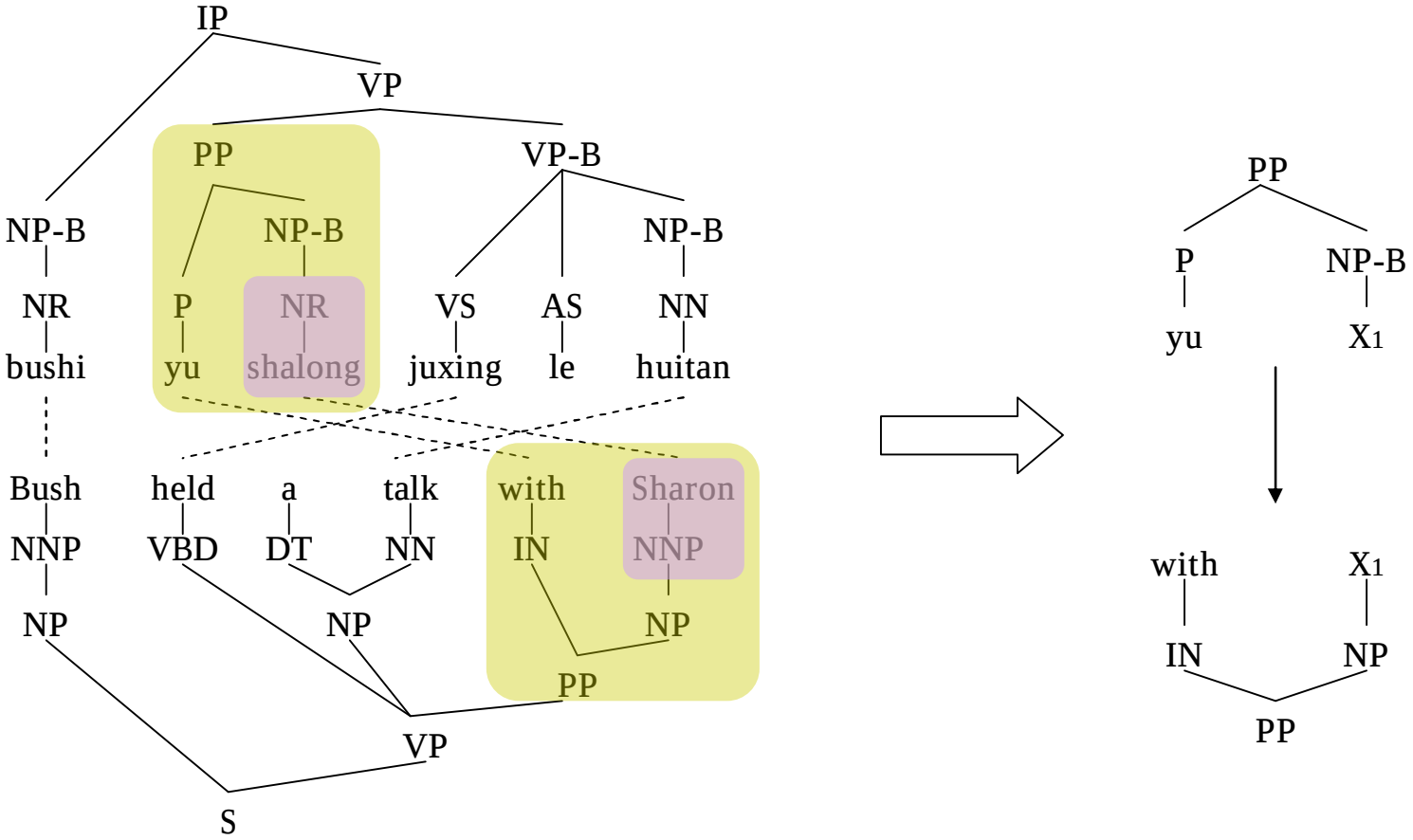
Tree-to-Tree Translation

- Tree-to-Tree (e.g., Eisner 2003, Zhang et al., 2008)



Tree-to-Tree Translation

- Tree-to-Tree (e.g., Eisner 2003, Zhang et al., 2008)

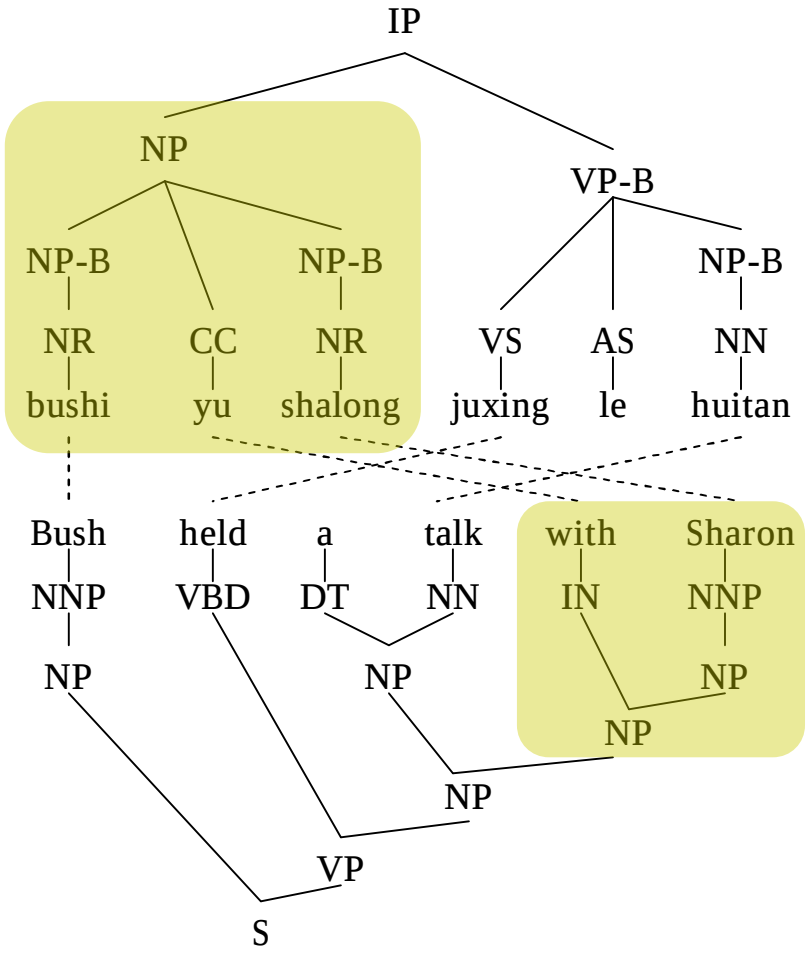


Challenges

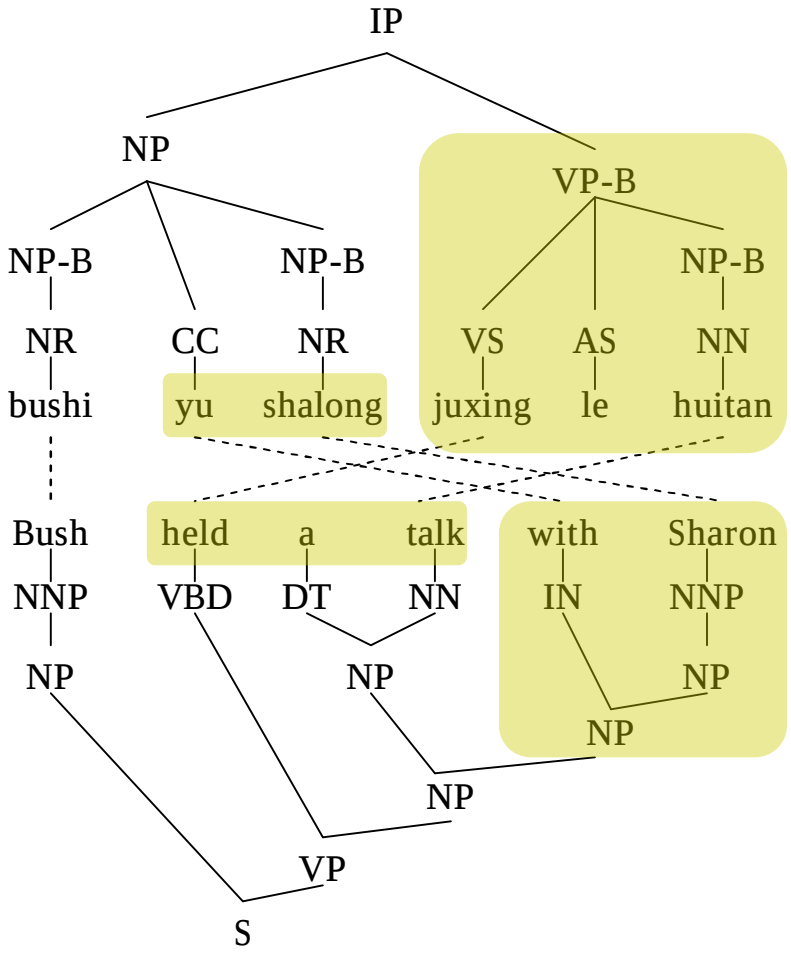
- Tree-to-tree approaches face two major challenges:
  - most vulnerable to parsing error
  - poorest rule coverage

Parsing Error

both trees can be ill-formed!



Rule Coverage



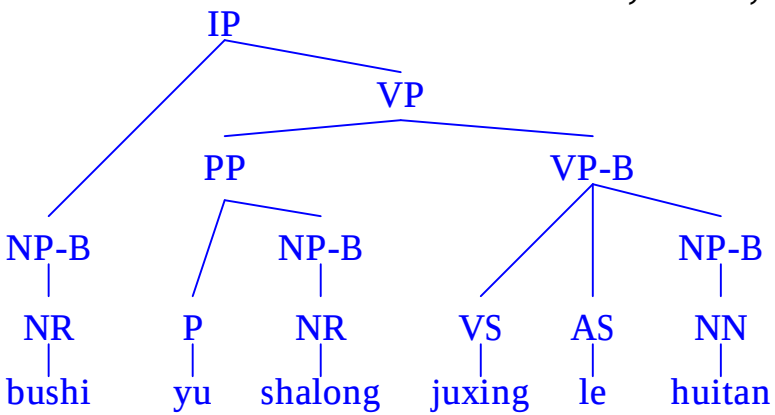
This Work

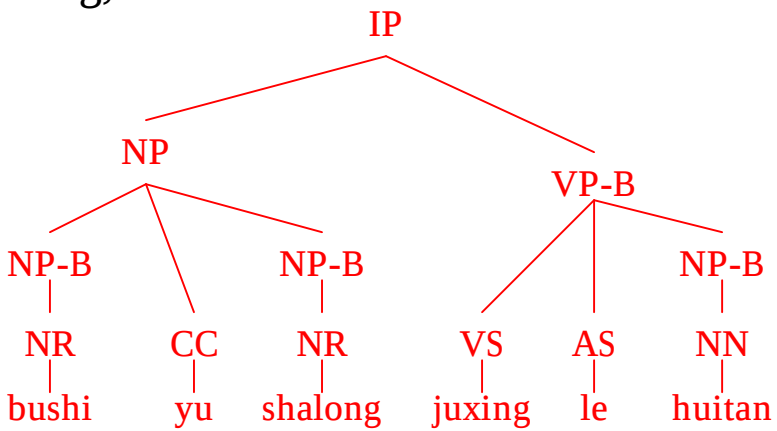
INSTITUTE OF COMPUTING TECHNOLOGY

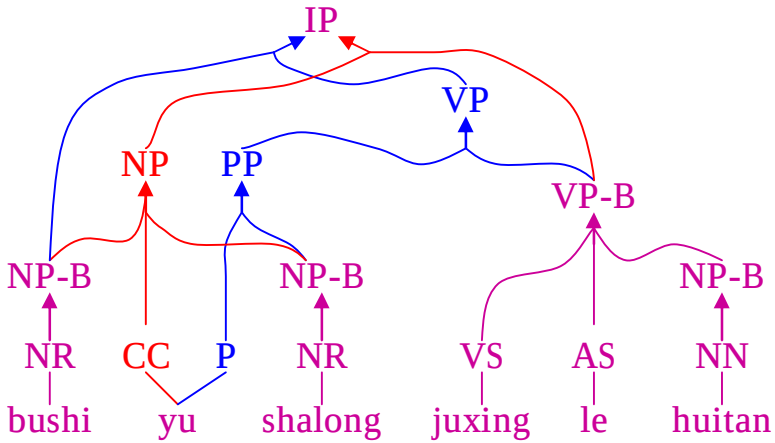
- We replace 1-best trees with packed forests to alleviate the two problems:
  - parsing error
  - rule coverage
- Our approach outperforms the tree-based system dramatically (+3.6) and achieves comparable performance with Moses.

Packed Forest

Tomita, 1985; Billot and Lang, 1989

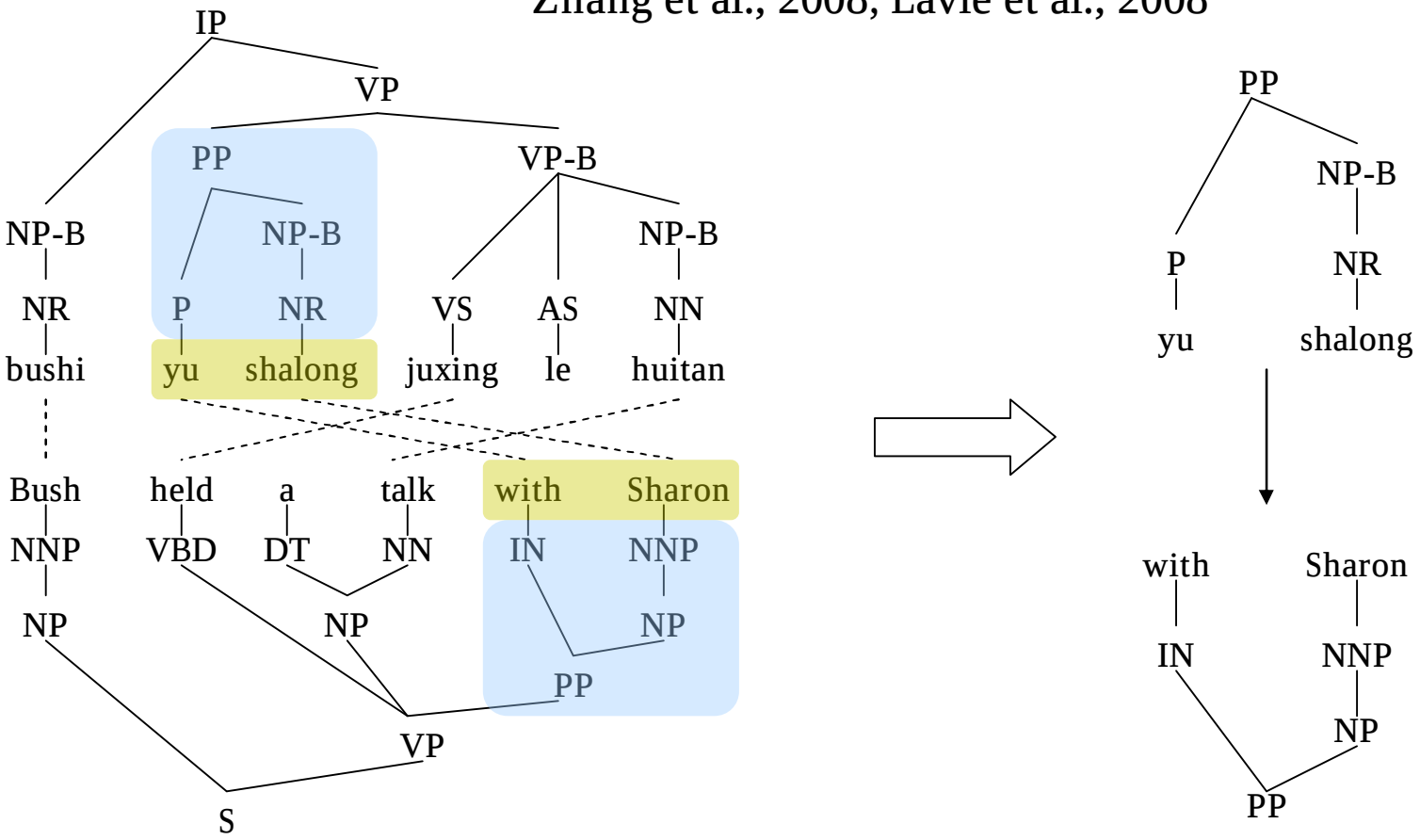




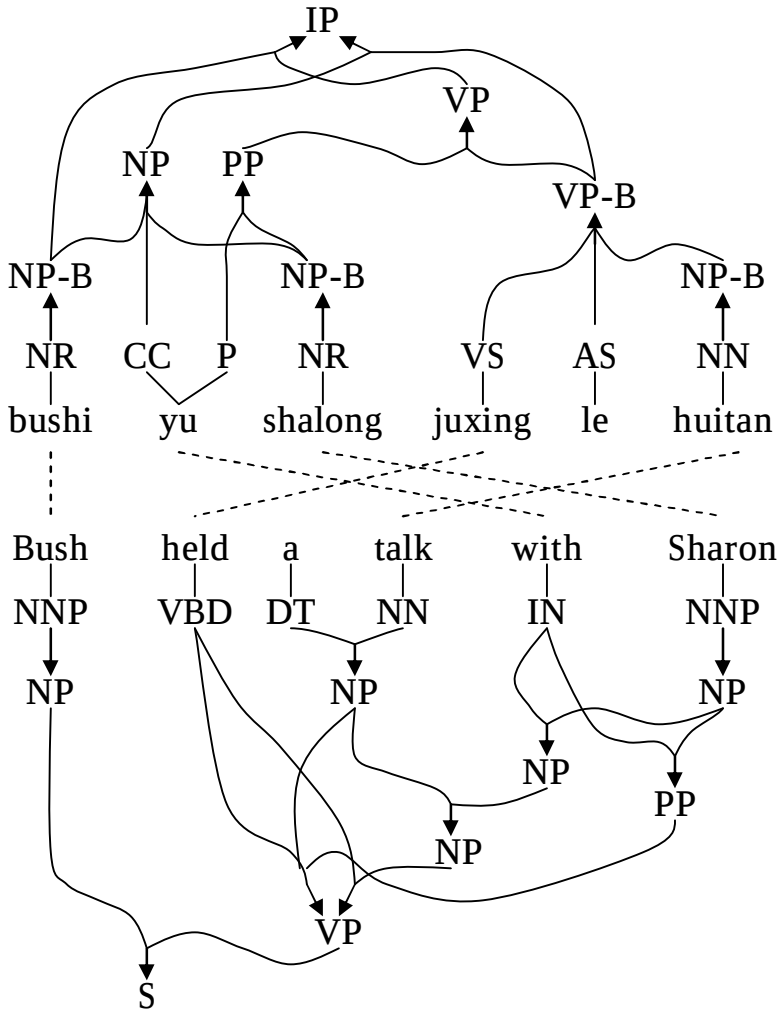


Tree-based Rule Extraction

Zhang et al., 2008; Lavie et al., 2008



Forest-based Rule Extraction

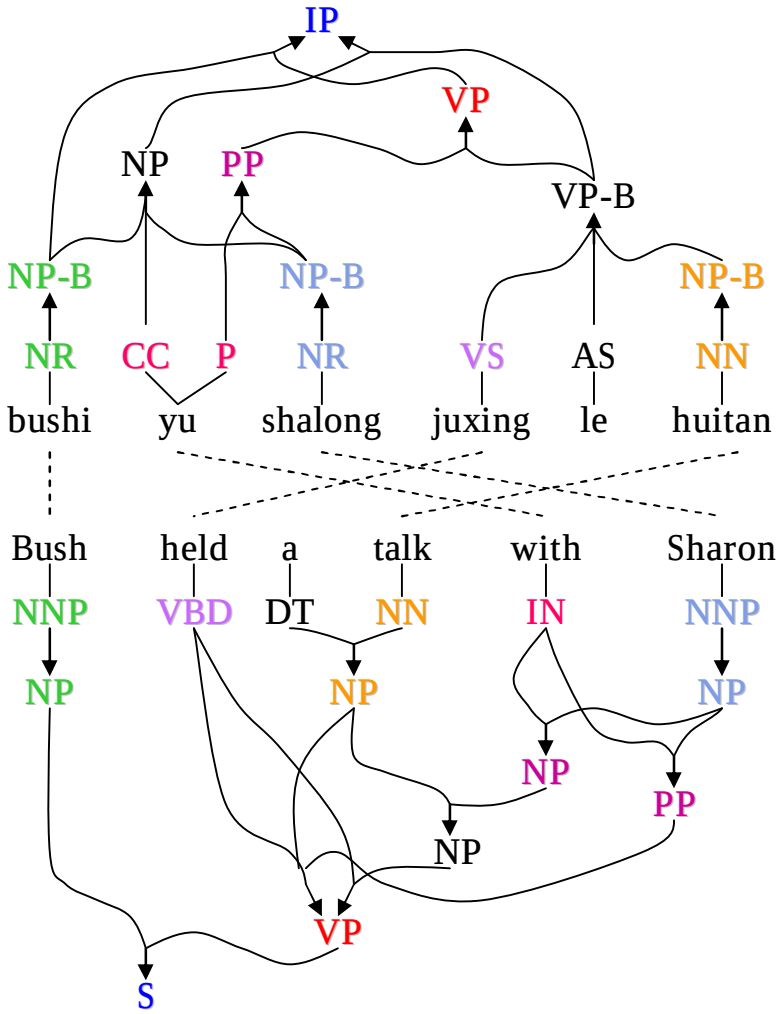


Forest-based Rule Extraction

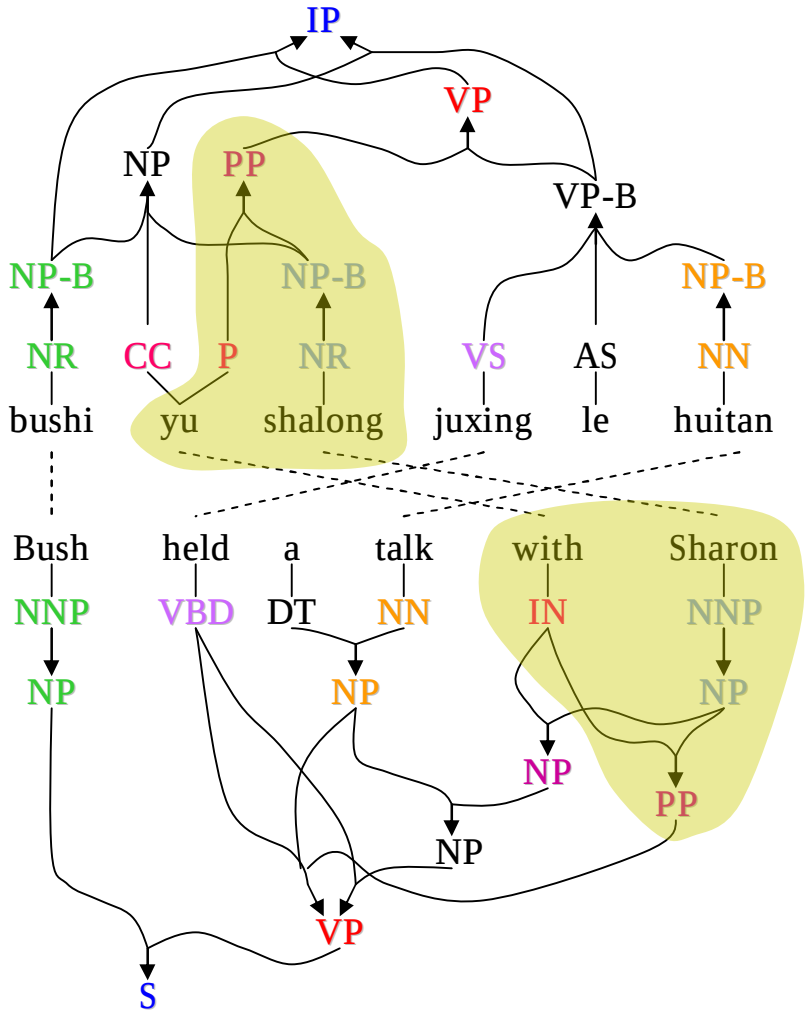
INSTITUTE OF COMPUTING TECHNOLOGY

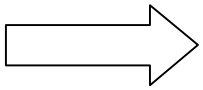
- Following GHKM (Galley et al., 2004), our extraction method invovles three steps:
  - identify the correspondence between nodes
  - identify minimal rules
  - get composed rules

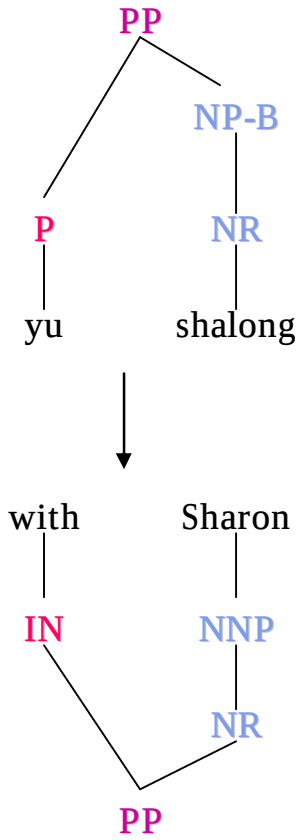
Frontier Nodes



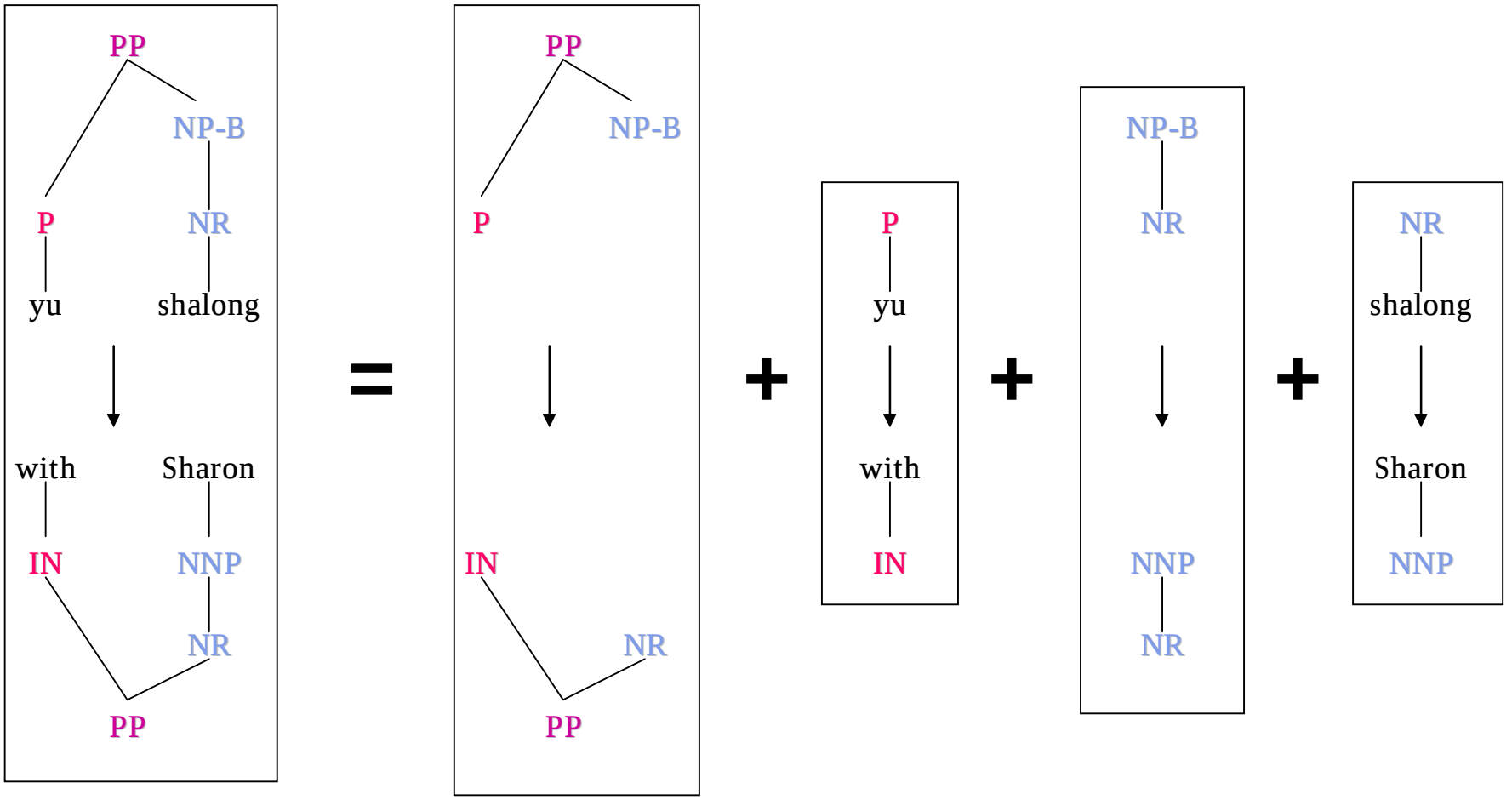
Identify Tree Pairs







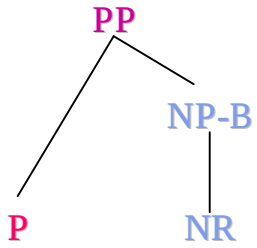
Decomposition



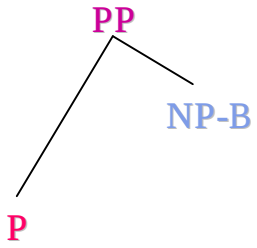
Frontier Trees and Tree Pairs

INSTITUTE OF COMPUTING TECHNOLOGY

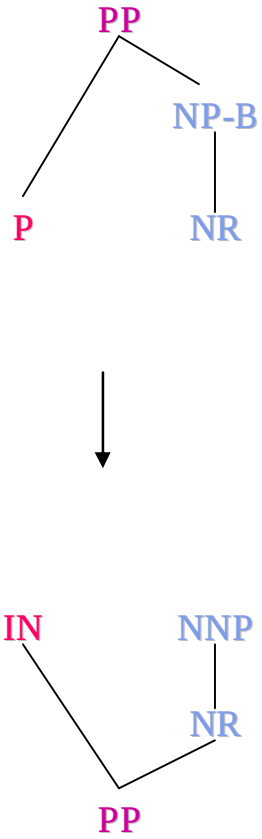
frontier tree



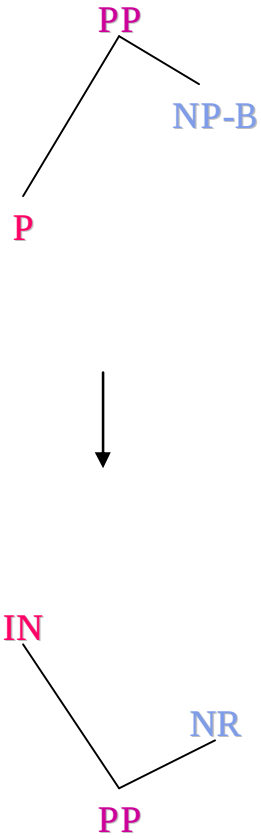
minimal frontier tree



frontier tree pair

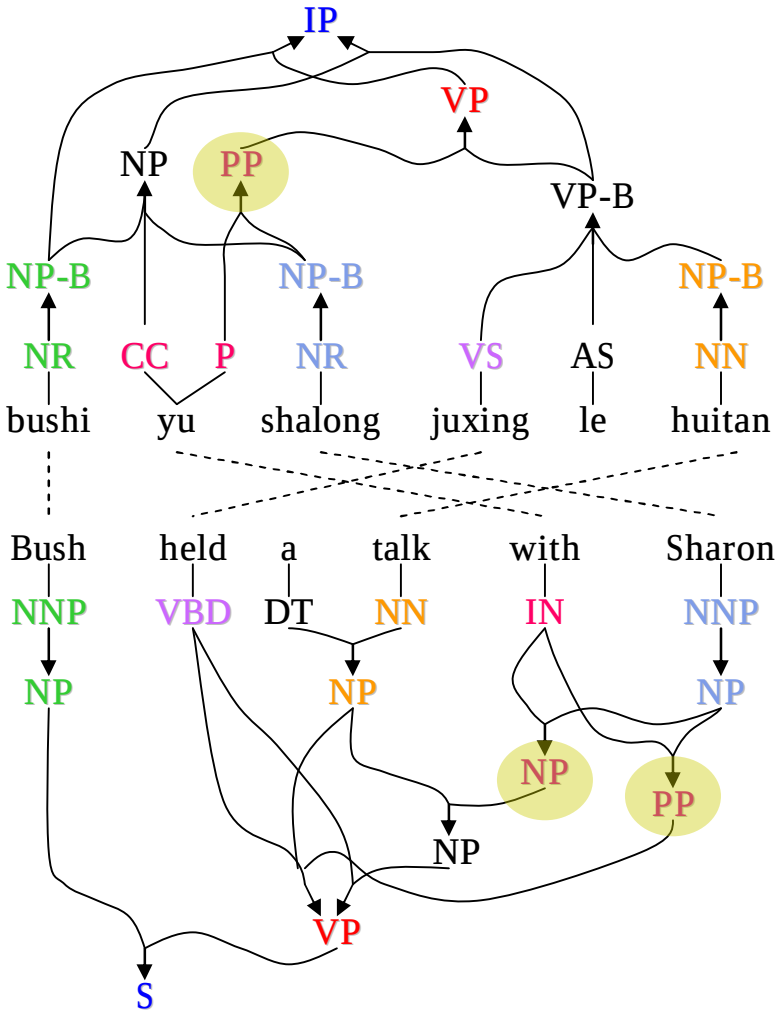


minimal frontier tree pair



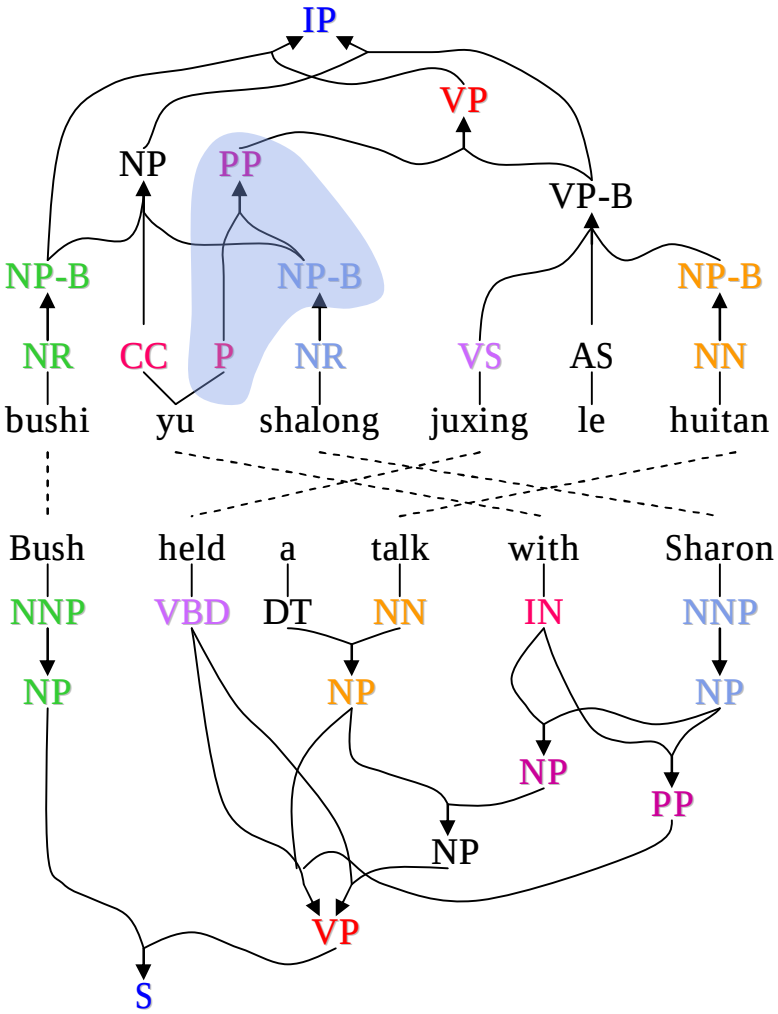
Identify Tree Pairs for the Node PP

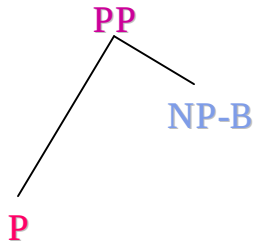
step 1: identify corresponding frontier nodes



Identify Tree Pairs for the Node PP

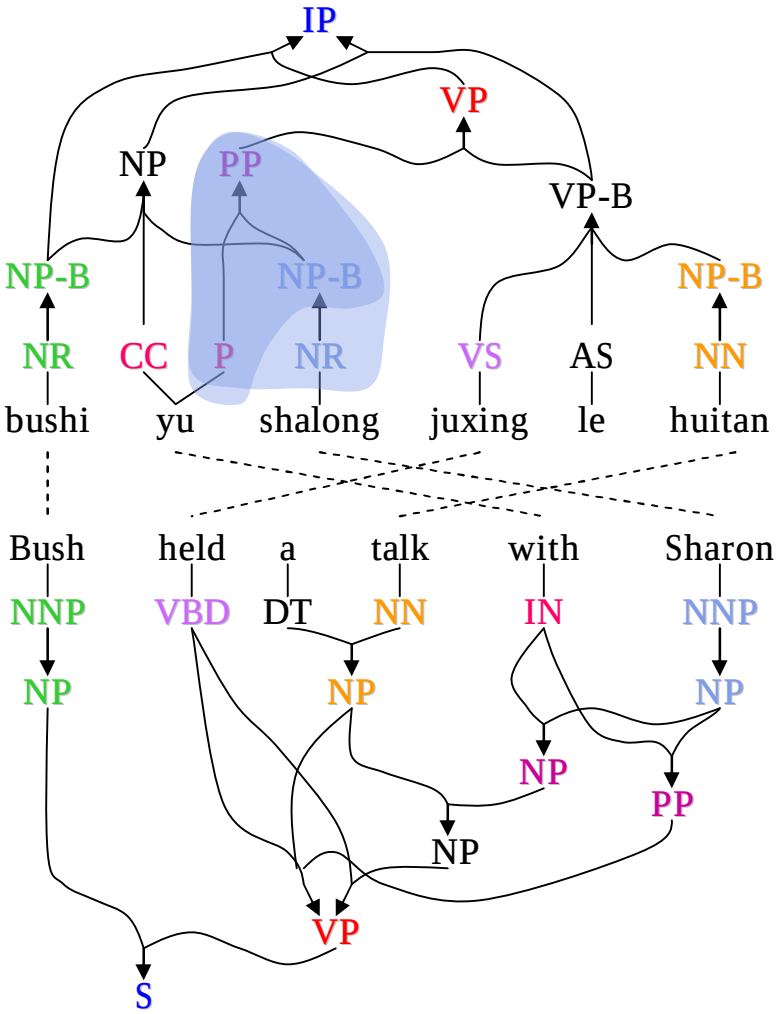
step 2: identify frontier trees for each node

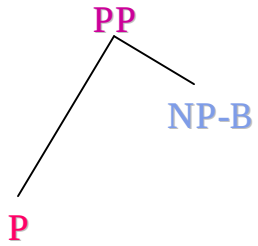


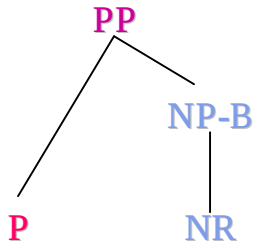


Identify Tree Pairs for the Node PP

step 2: identify frontier trees for each node

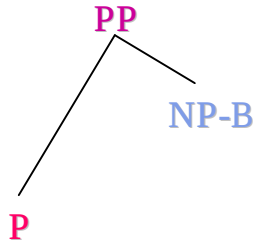


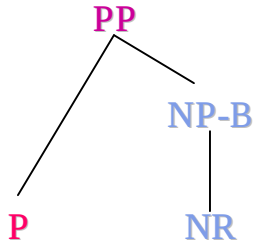


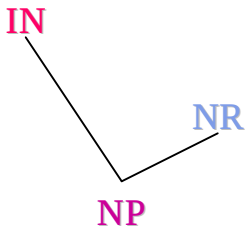


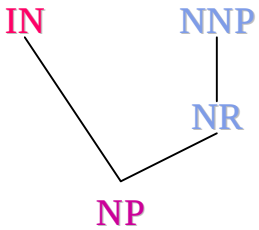
Identify Tree Pairs for the Node PP

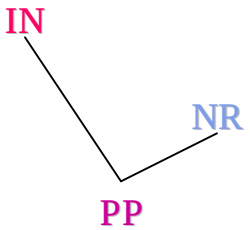
step 2: identify frontier trees for each node

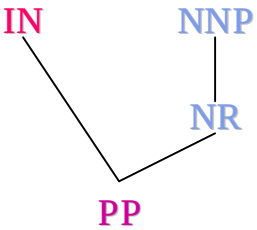






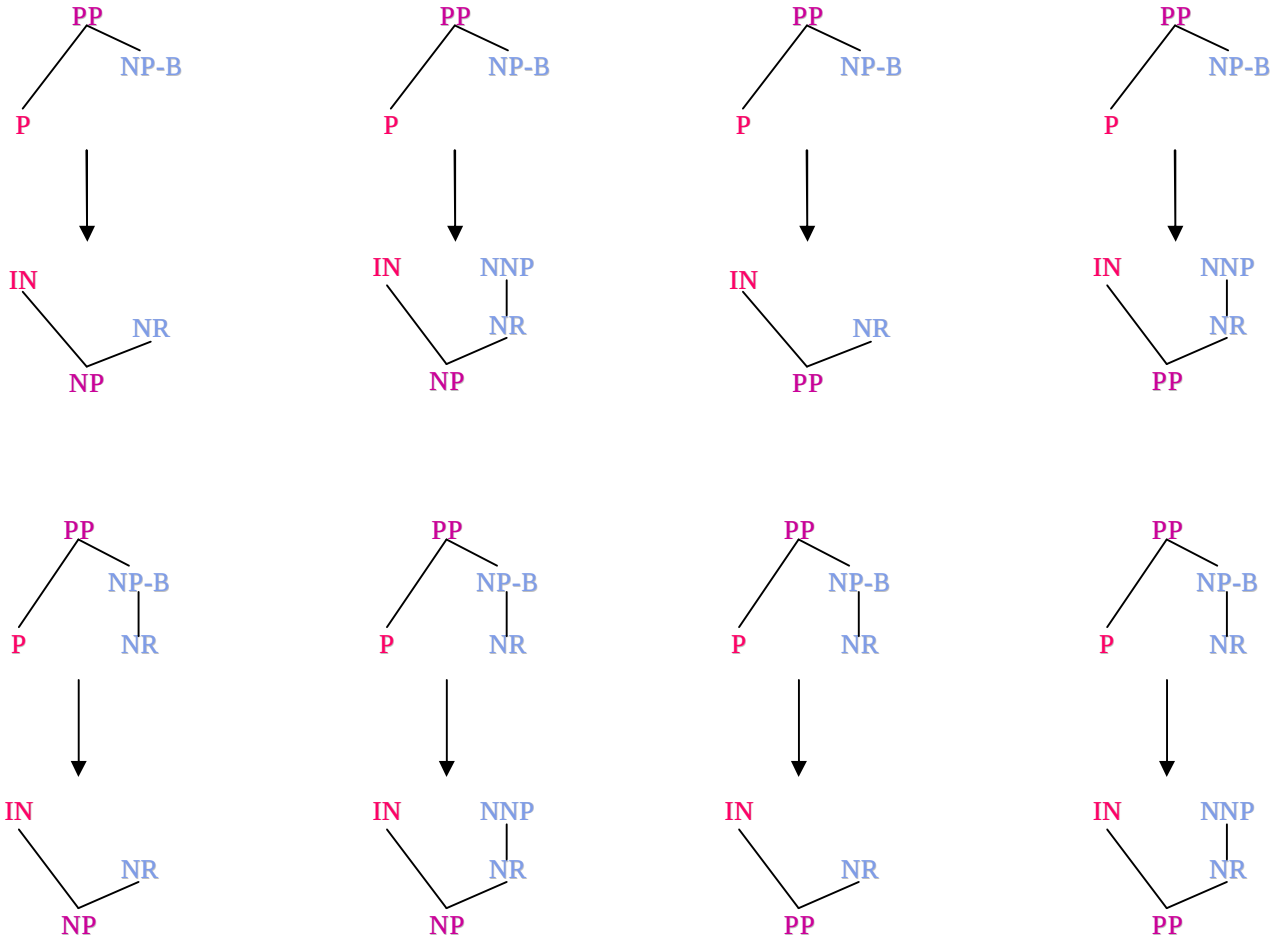






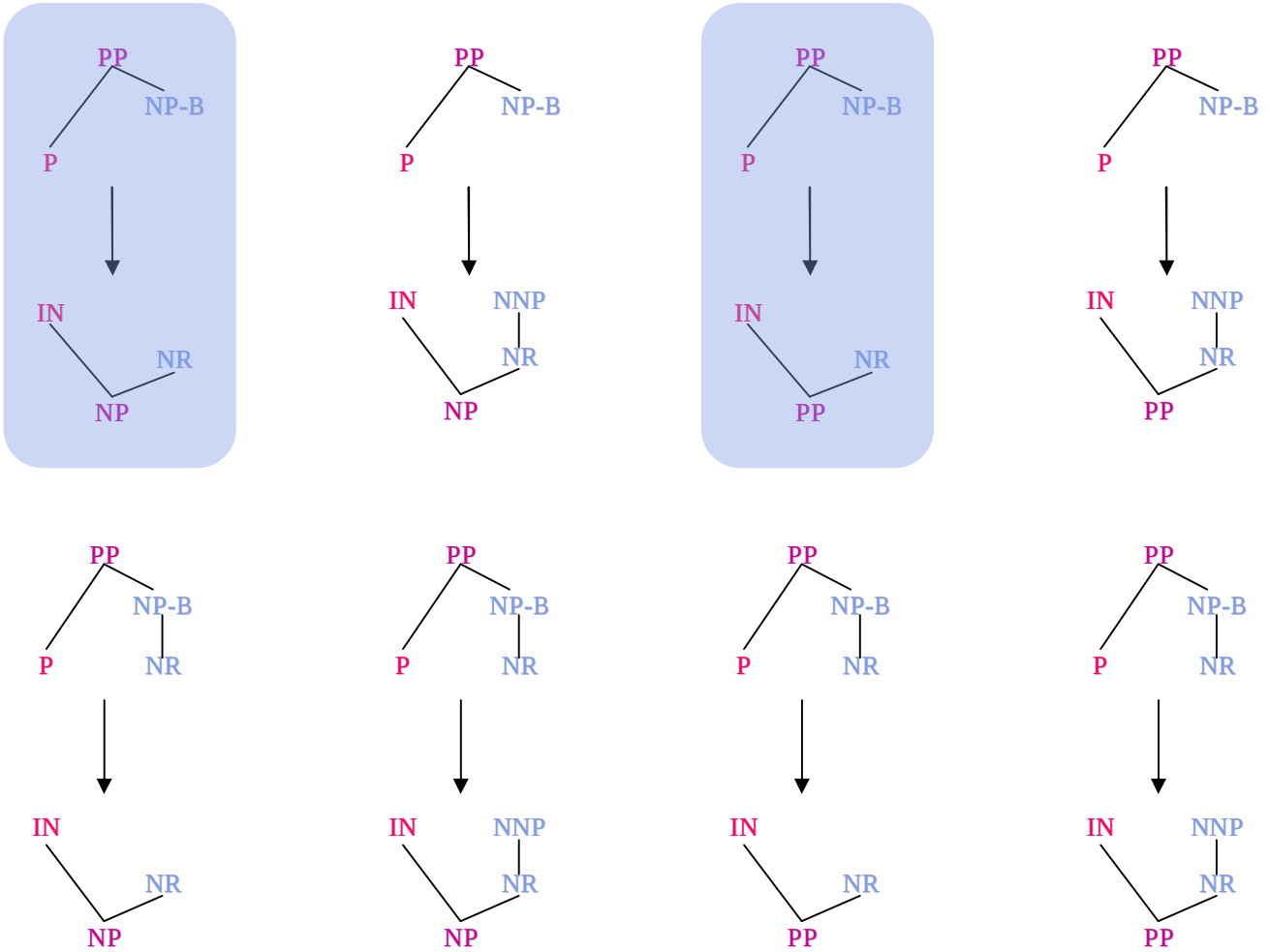
Identify Tree Pairs for the Node PP

step 3: build frontier trees pairs



Identify Tree Pairs for the Node PP

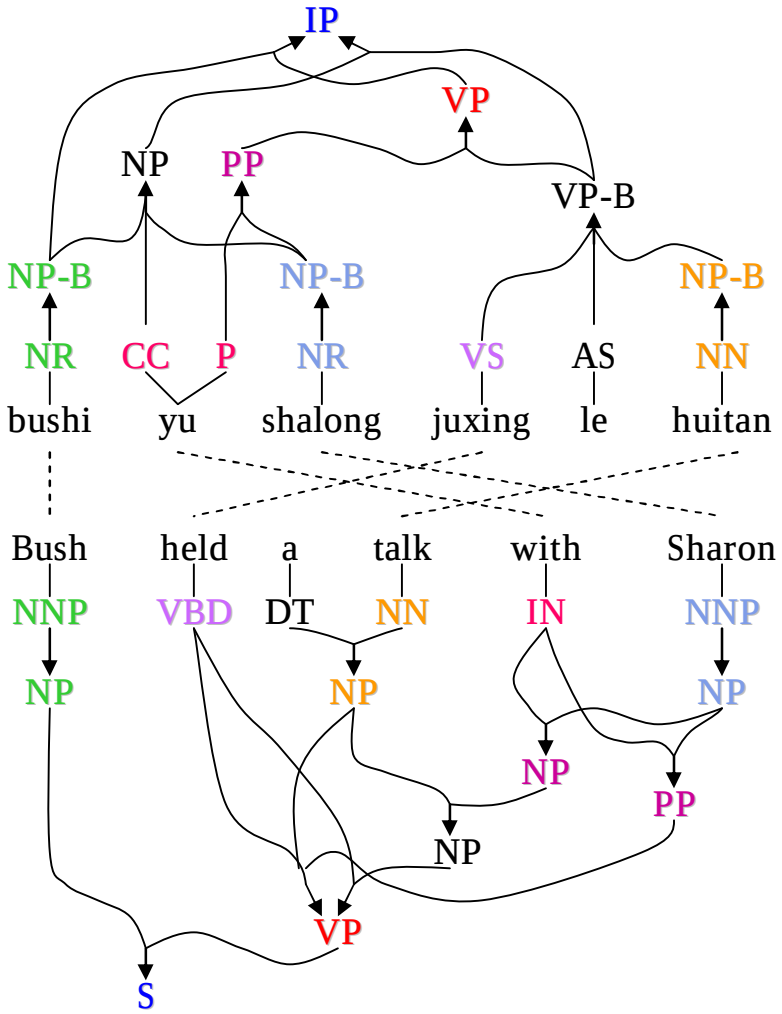
step 4: identify minimal frontier trees pairs



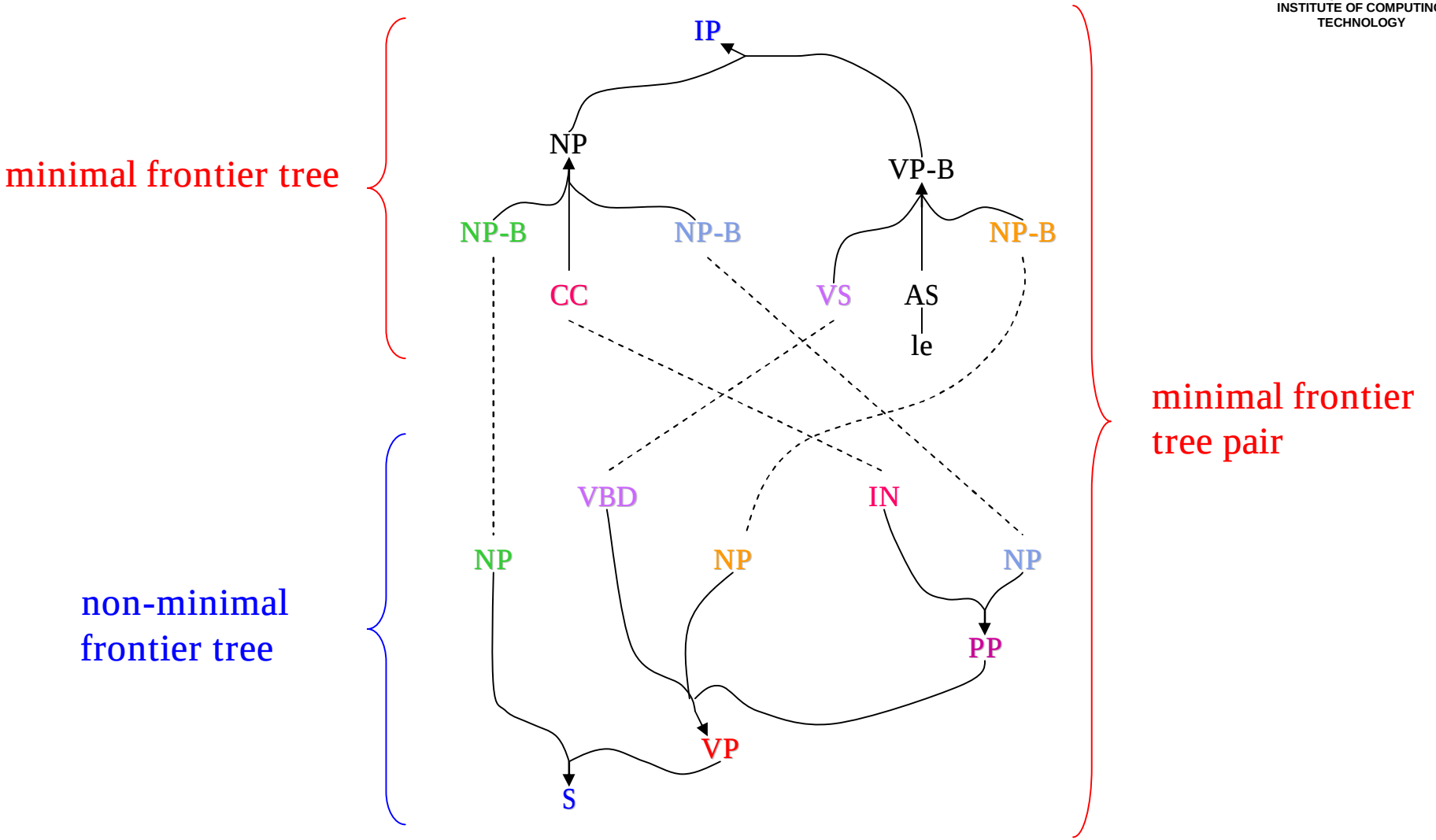
Difficulty in Finding Tree Pairs

A minimal frontier tree pair is not necessarily a pair of minimal frontier trees

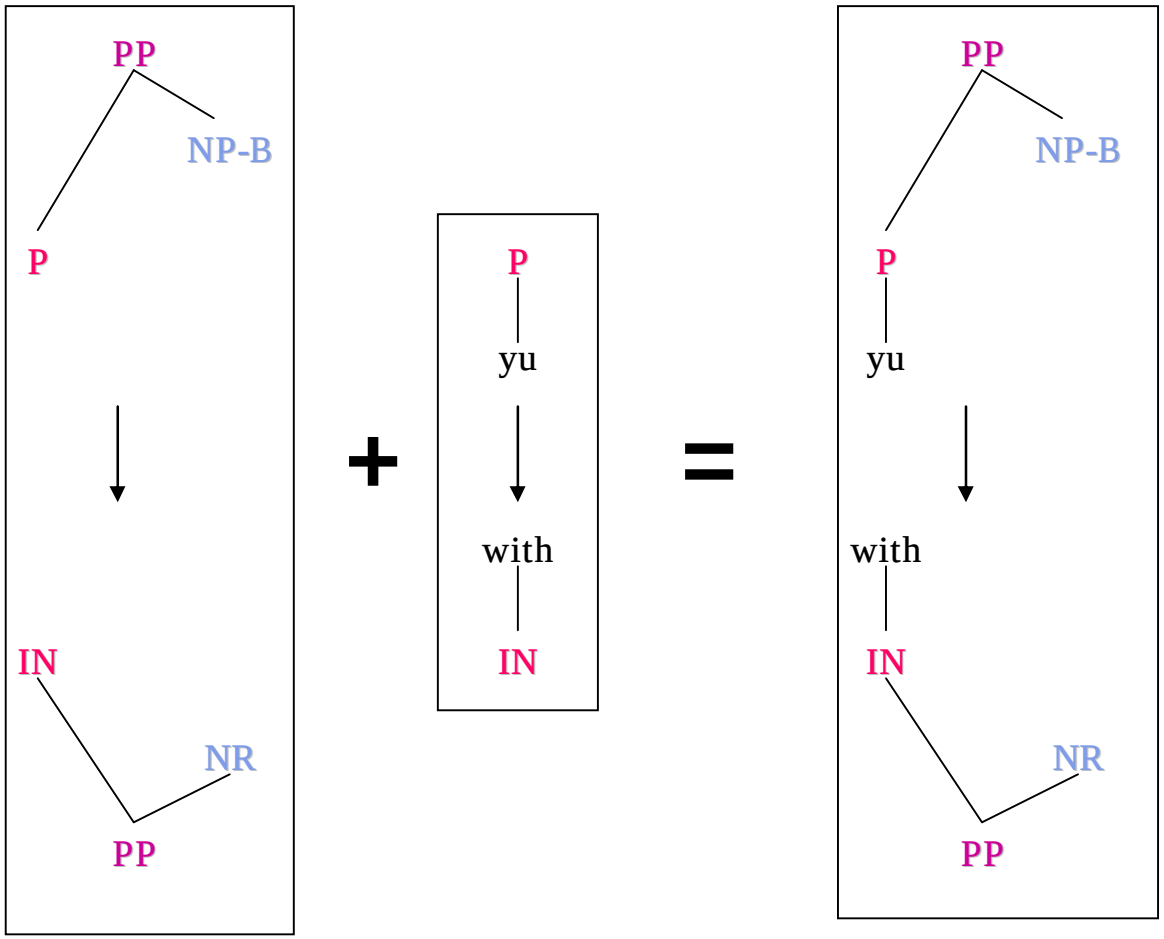
Difficulty in Finding Tree Pairs



Difficulty in Finding Tree Pairs



Get Composed Rules



Experiments

INSTITUTE OF COMPUTING TECHNOLOGY

- Chinese-to-English translation
- Training set: 31K sentence pairs with 840K Chinese words and 950K English words
- Language Model: 3-gram trained on the English side
- Development set: NIST 2002
- Test set: NIST 2005
- Metric: case-insensitive BLEU4

Tree-based Vs. Forest-based

INSTITUTE OF COMPUTING TECHNOLOGY

p	avg. trees	# of rules	BLEU4
0	1	74K	20.21
2	238.94	105K	21.65
5	5.78M	348K	23.36
8	65.9M	574K	23.73
10	105M	743K	23.85

Extraction and Decoding Time

INSTITUTE OF COMPUTING TECHNOLOGY

p	avg. trees	extraction	decoding
0	1	1.26	6.76
2	238.94	2.35	8.52
5	5.78M	6.34	14.87
8	65.9M	8.51	19.78
10	105M	10.21	25.81

extraction: milliseconds / sentence pair

decoding: seconds / sentence

Comparison with Moses

INSTITUTE OF COMPUTING TECHNOLOGY

training	Moses	this work
840K+950K	23.66	23.85
7.39M+9.41M	30.43	30.59

Conclusion

- Packed forests help alleviate two problems that tree-to-tree approaches face:
  - negative impact of parsing mistakes on translation quality
  - poor rule coverage

Thanks!