# 树到串模型
# 规则表示的一点思考


智能信息重点实验室自然语言处理组
谢军 吕雅娟 刘群

# 大纲

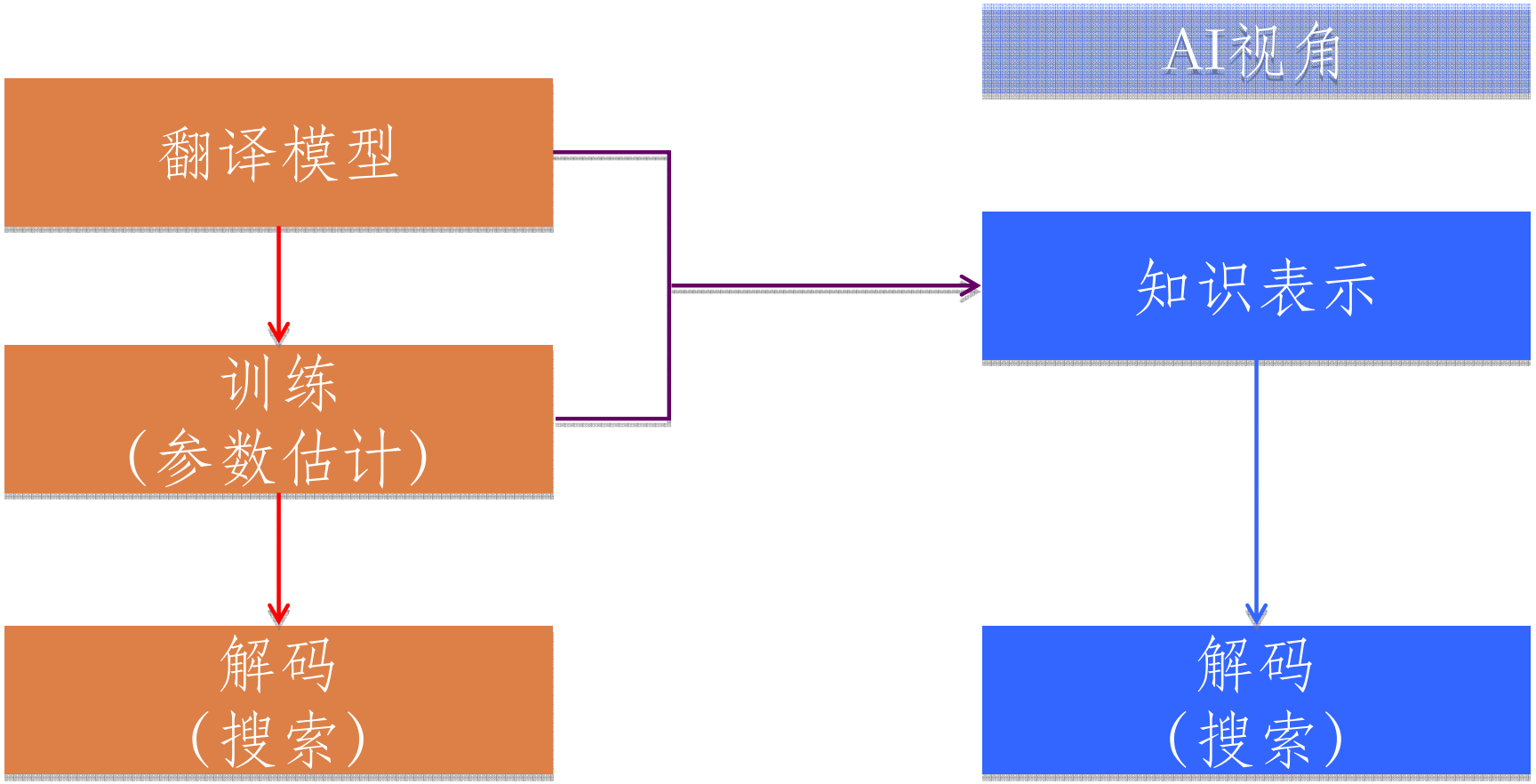
- 缘起
- 树到串模型的规则表示
  - 成分树到串模型
  - 依存树到串模型
- 关于规则表示的一点思考
  - 需要解决的问题
  - 现有规则表示的不足
- 总结

# 大纲

- 缘起
- 树到串模型的规则表示
  - 成分树到串模型
  - 依存树到串模型
- 关于规则表示的一点思考
  - 需要解决的问题
  - 现有规则表示的不足
- 总结

# 缘起

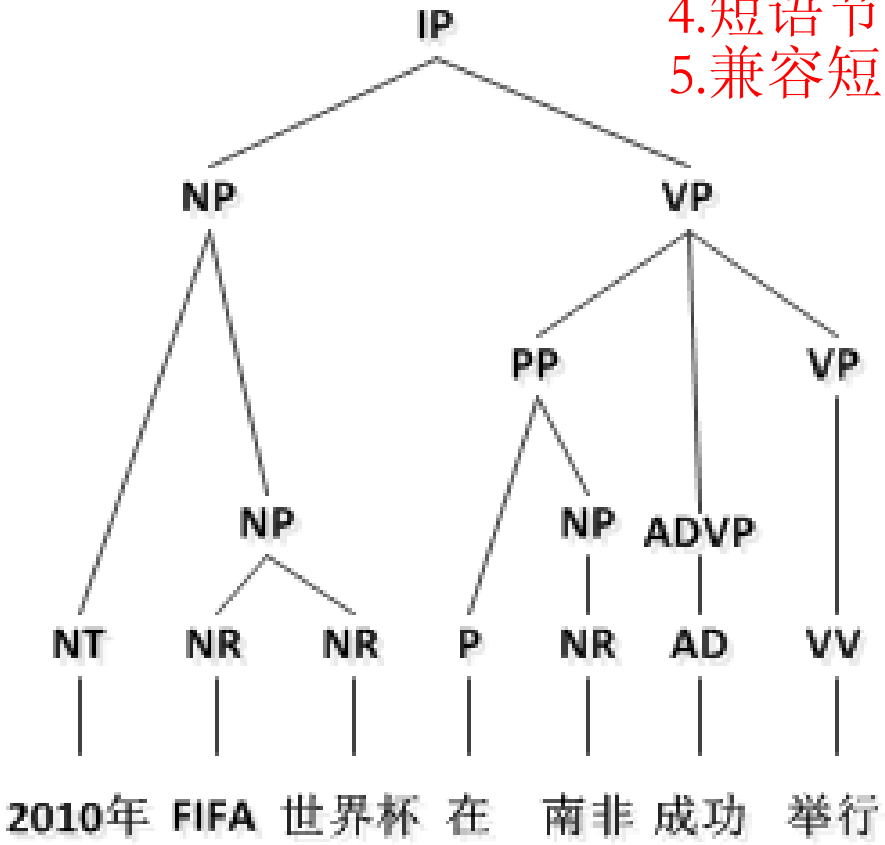
- SMT的三大任务



# 缘起

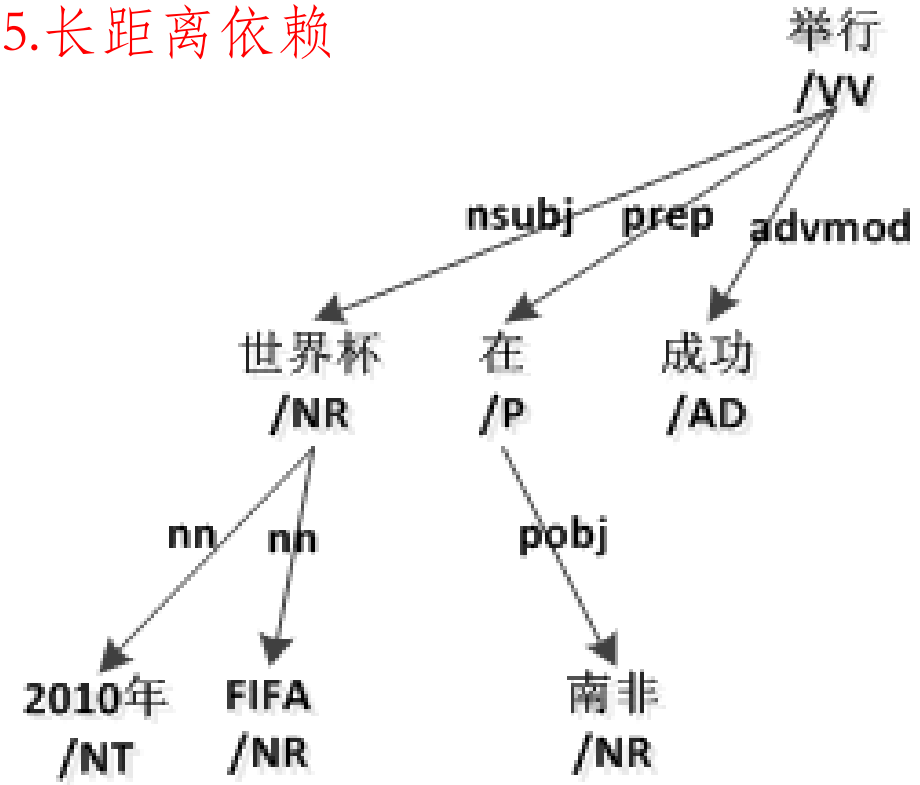
成分树

1.生成
2.文法
3.近邻组合
4.短语节点
5.兼容短语




1.分析
2.无文法
3.修饰关系
4.词汇化
5.长距离依赖

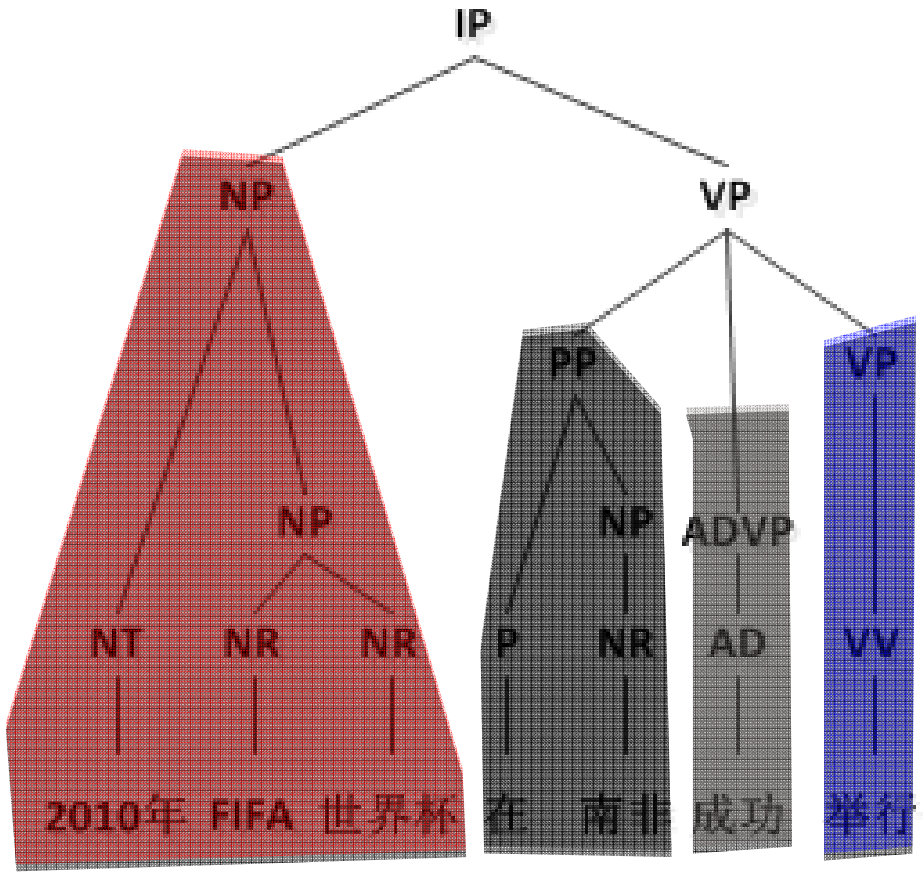
依存树

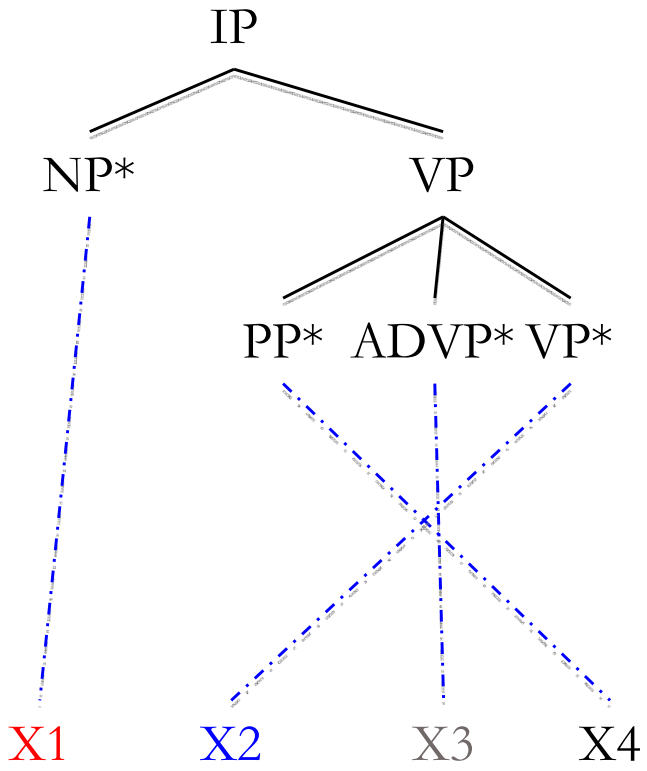


# 大纲

- 缘起
- 树到串模型的规则表示
  - 成分树到串模型
  - 依存树到串模型
- 关于规则表示的一点思考
  - 需要解决的问题
  - 现有规则表示的不足
- 总结

# 成分树到串模型（Cons2Str）

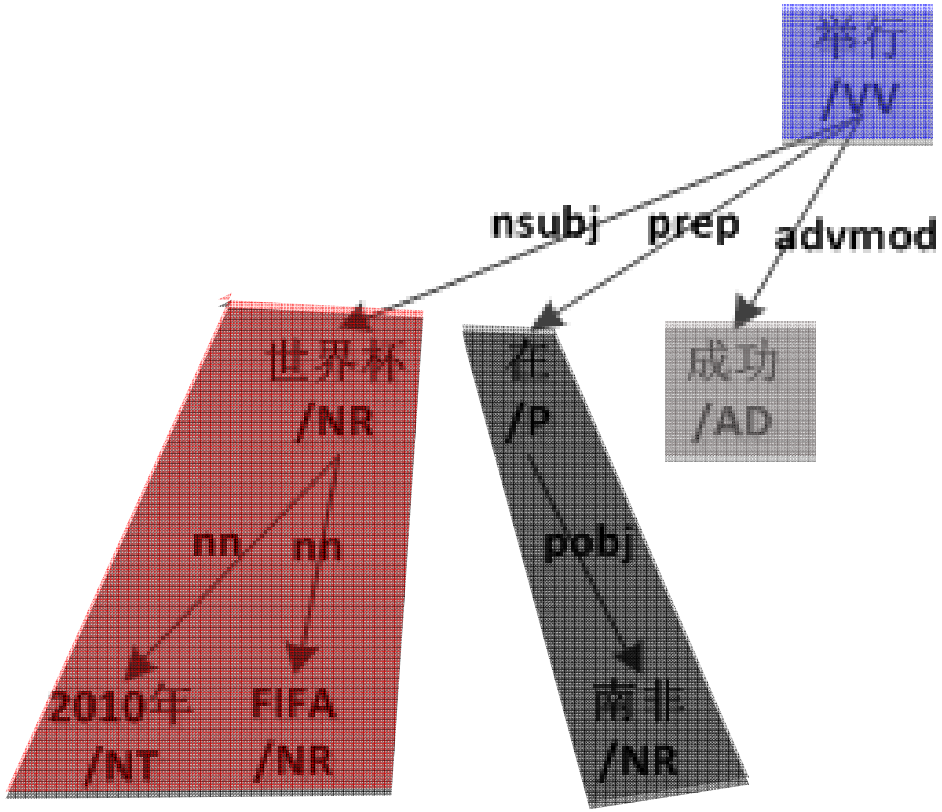


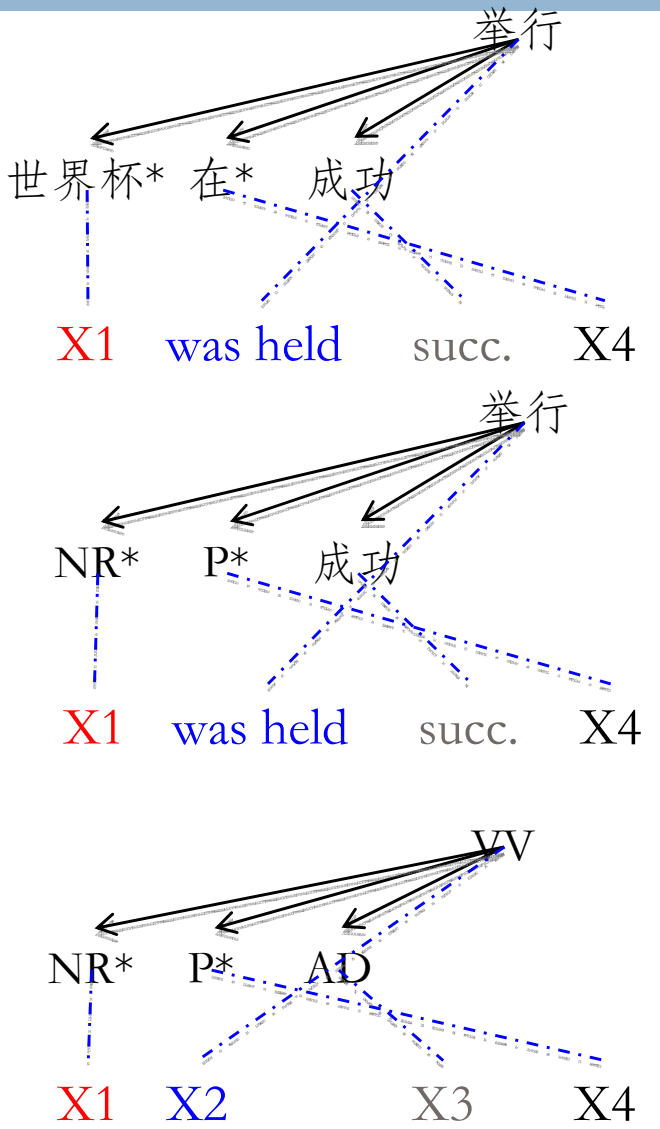

2010 FIFA World Cup was held successfully in South Africa

Liu et. al, 2006; Mi et.al, 2008

# 依存树到串模型（Dep2Str）




2010 FIFA World Cup was held successfully in South Africa




Xie et.al, 2011

# 大纲

- 缘起
- 树到串模型的规则表示
  - 成分树到串模型
  - 依存树到串模型
- 关于规则表示的一点思考
  - 需要解决的问题
  - 现有规则表示的不足
- 总结

# 需要解决的问题

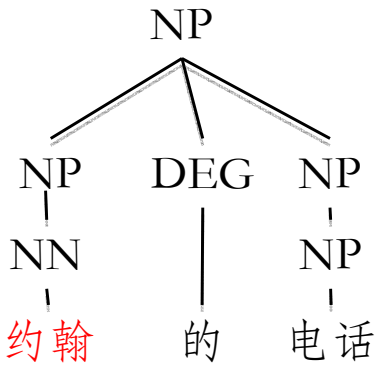
- 调序 ✔
- 时态（现在、过去、将来） ✗
- 语态（主动、被动） ✗
- 非组合短语 ✗

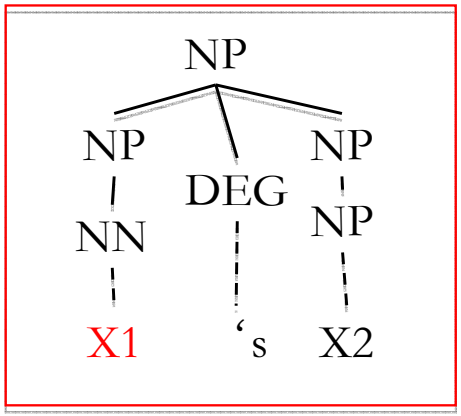
# 现有规则表示的不足

- 规则区分度差
  - 调序
  - 时态
  - 语态
- 短语兼容性差（Dep2Str）

# 现有规则表示的不足--调序

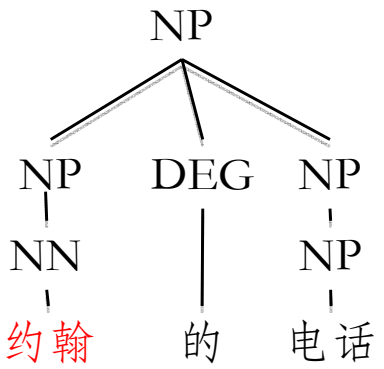
- 所有格
  - ‘s 所有格
    - 表示来源 eg. John ‘s telephone
  - Of所有格
    - 中心词为无生命的事物 eg. the subject of the sentence

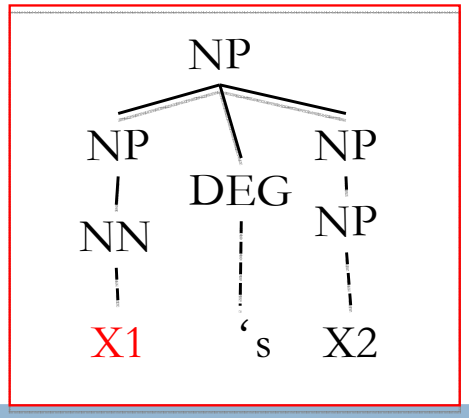


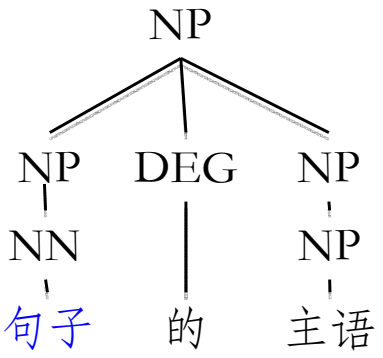



# 现有规则表示的不足--调序

- 所有格
  - ‘s 所有格
    - 表示来源 eg. John ‘s telephone
  - Of所有格
    - 中心词为无生命的事物 eg. the subject of the sentence

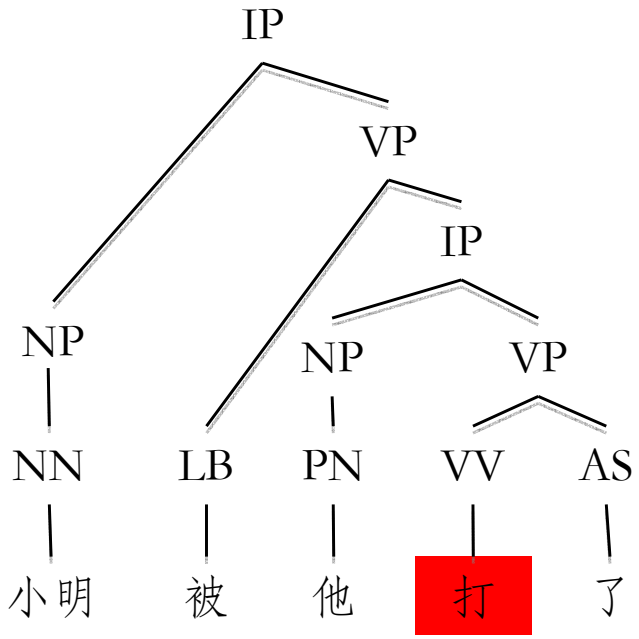


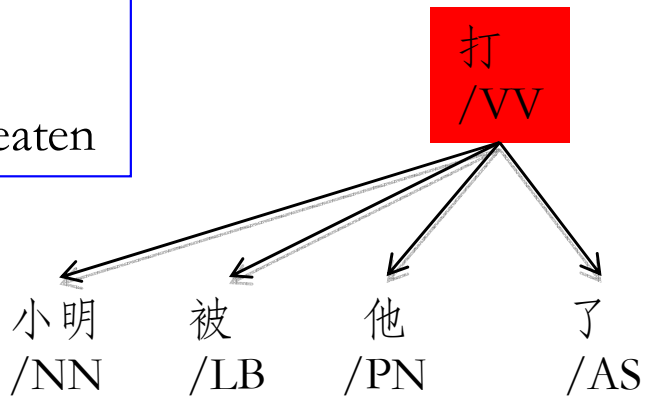




词义信息

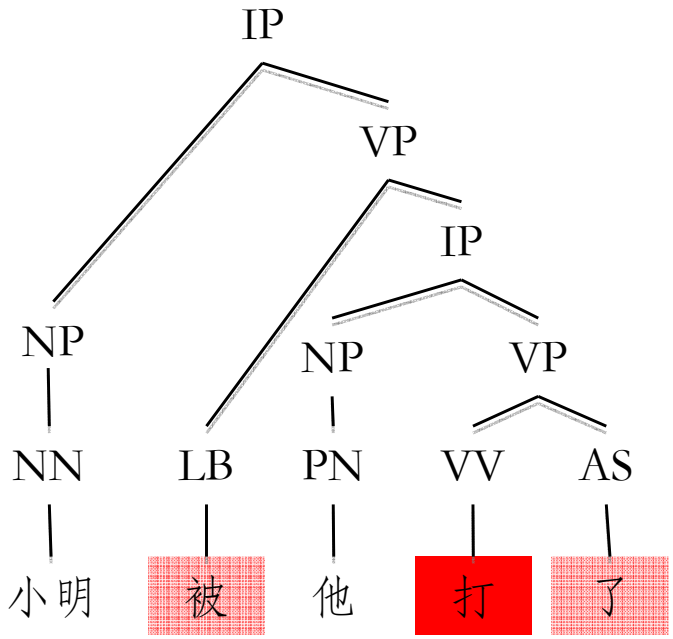
# 现有规则表示的不足—时态&语态



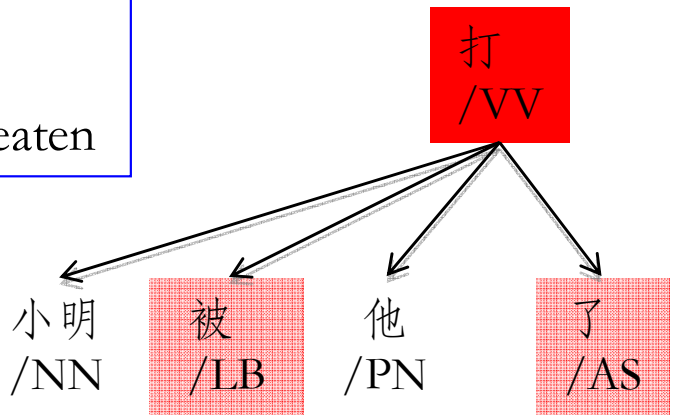

beat
was beaten
beaten
have been beaten



# 现有规则表示的不足—时态&语态



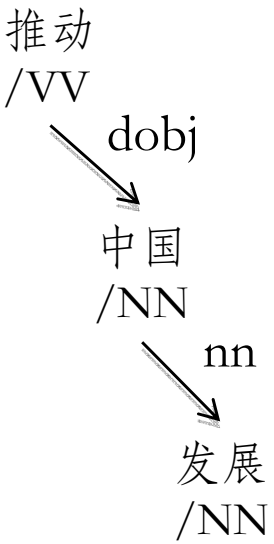

beat
was beaten
beaten
have been beaten

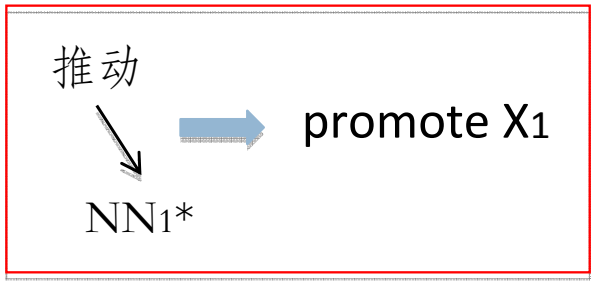


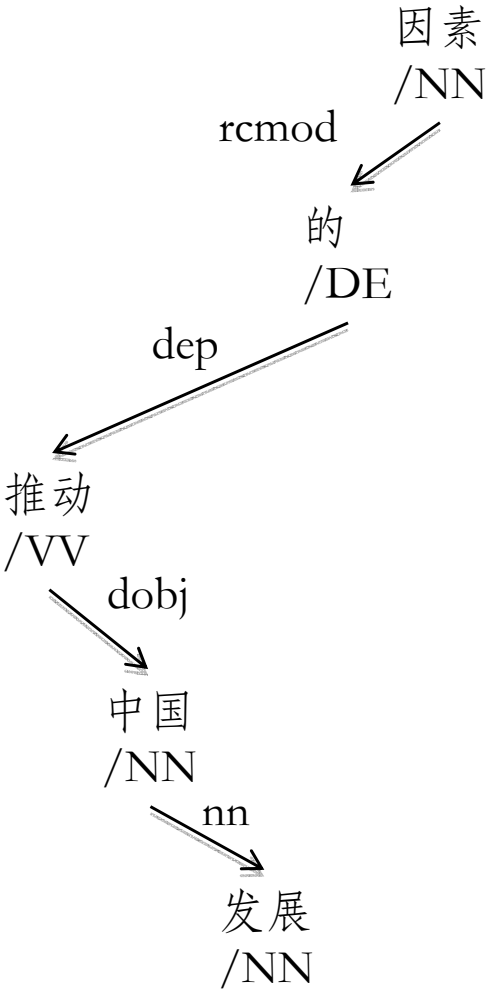

词汇化上下文信息
（被、把、了、着、过……）

# 现有规则表示的不足—时态&语态

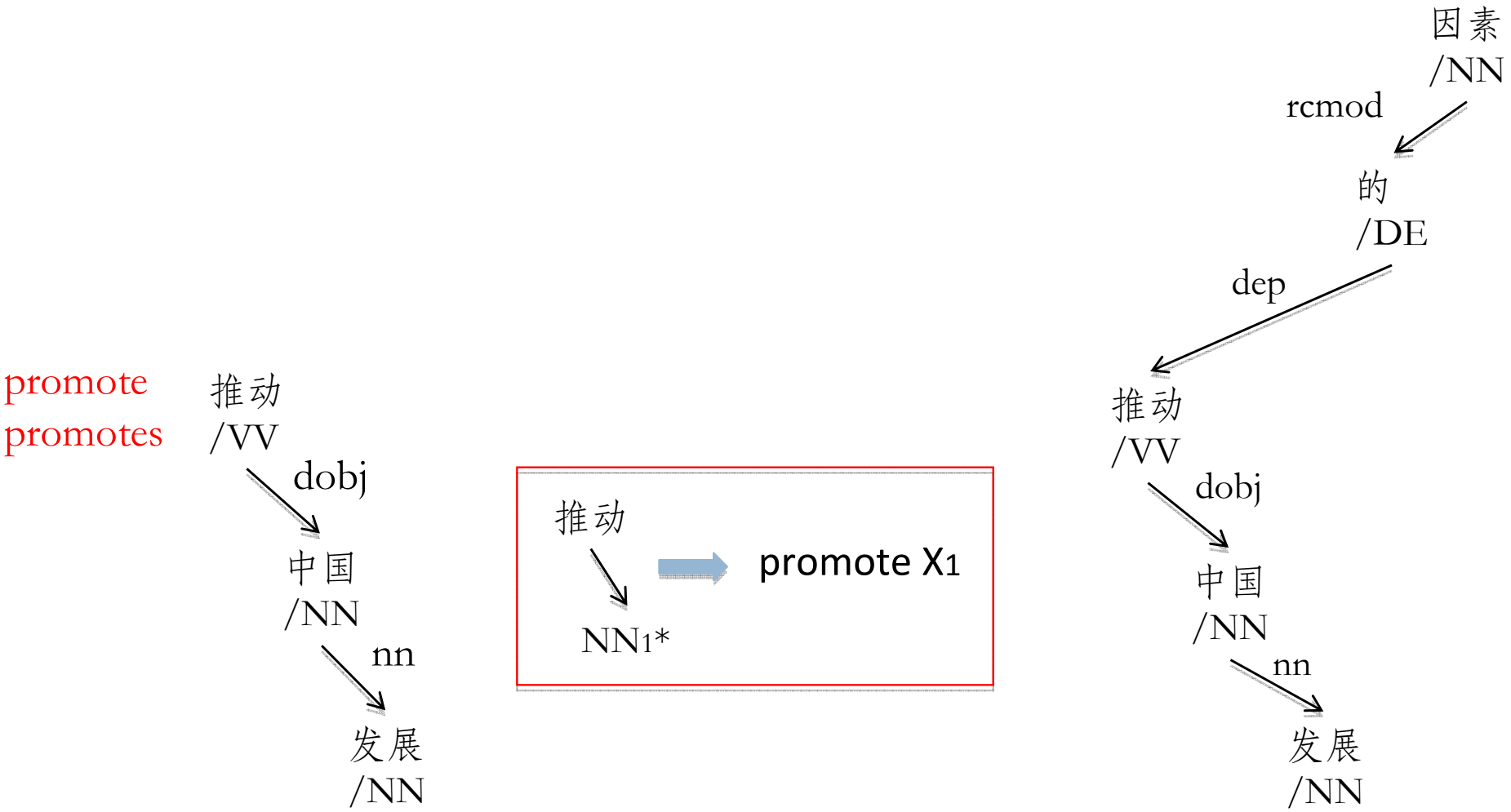
promote
promotes



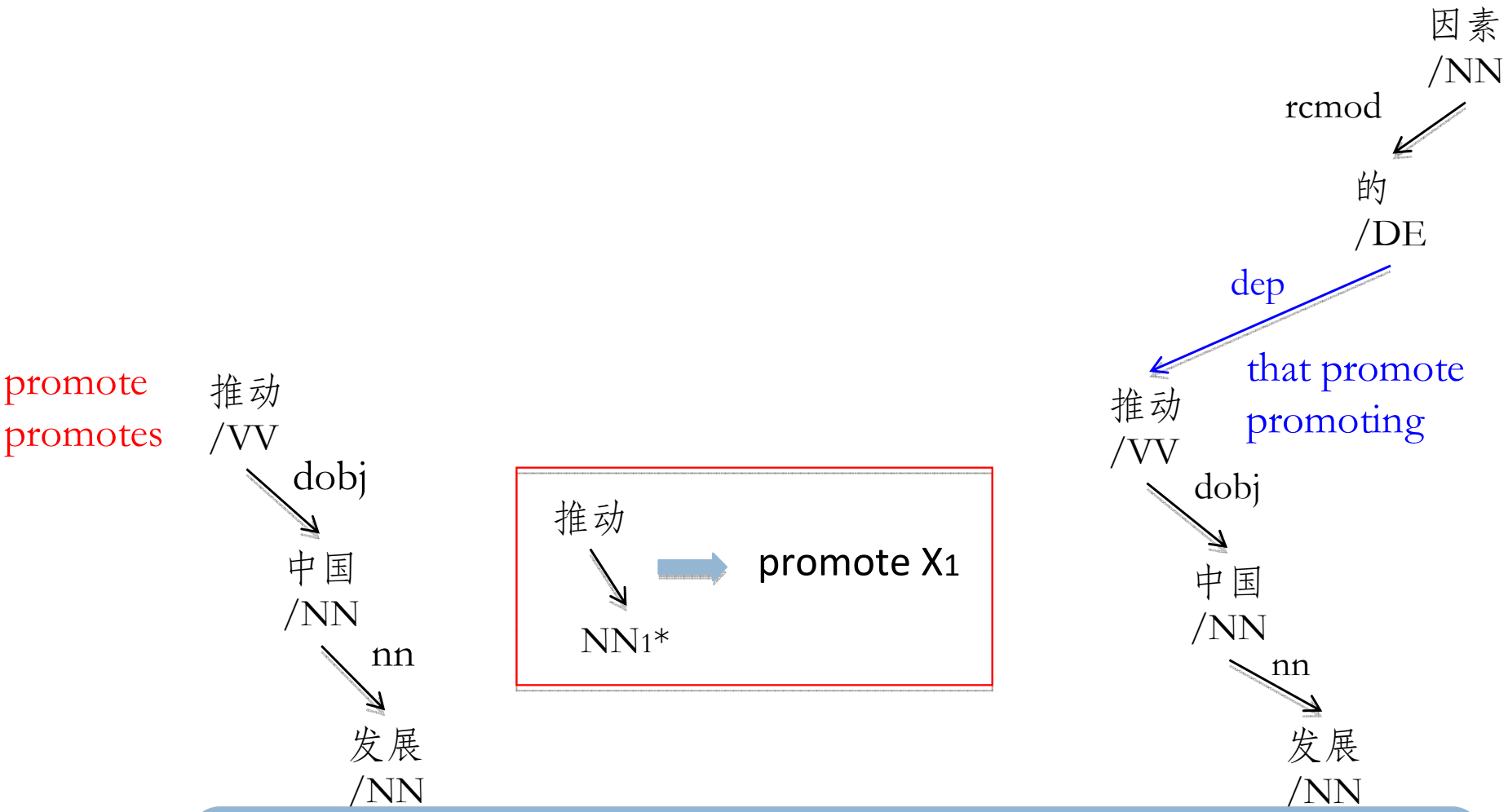



# 现有规则表示的不足—时态&语态



# 现有规则表示的不足—时态&语态



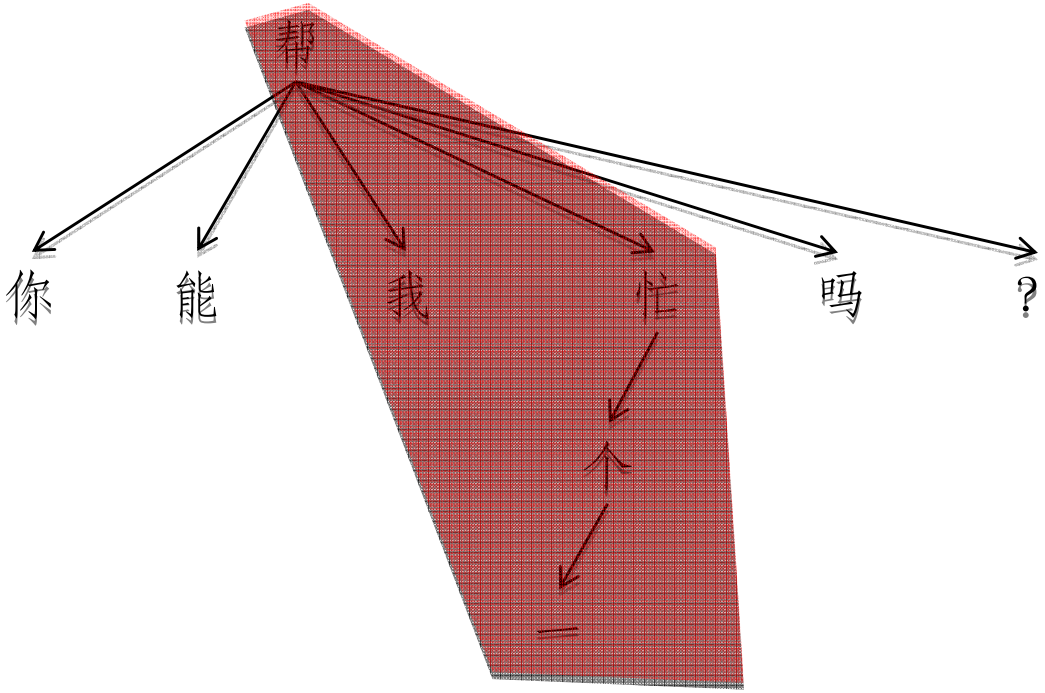
# 现有规则表示的不足
## ——短语兼容性（Dep2Str）

- 习惯用语

你 能 帮 我 一 个 忙 吗？

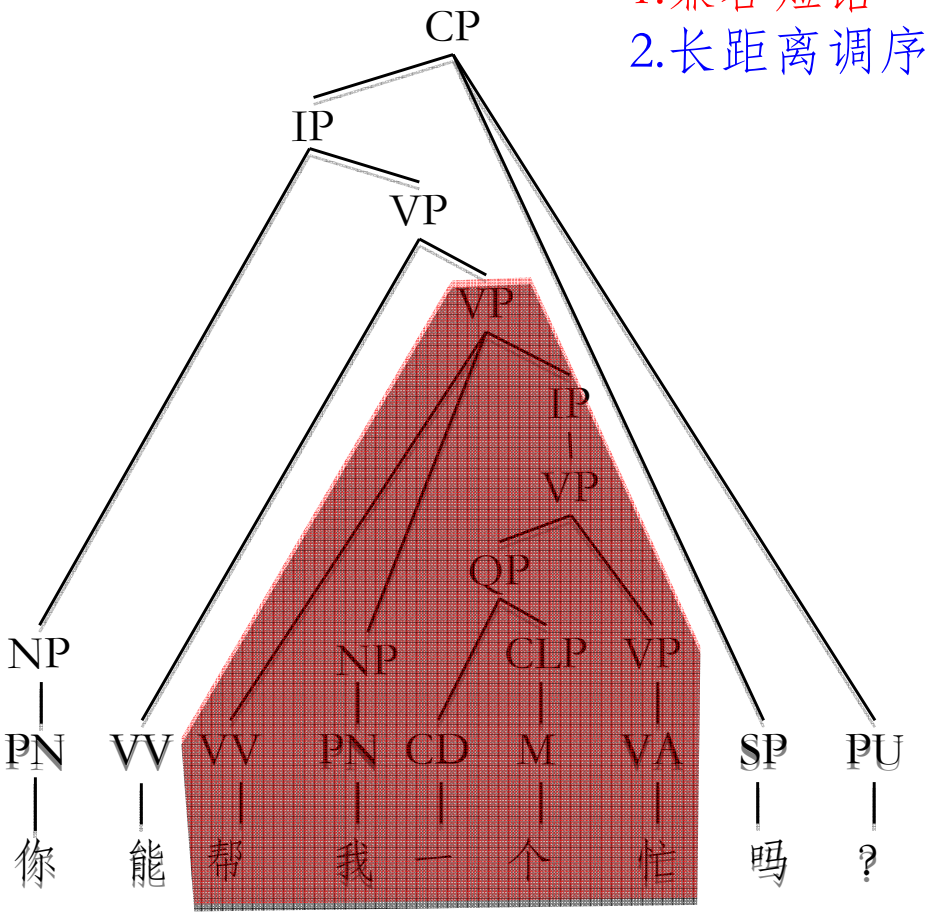
Could you do me a favor?

- HPB：X→(帮 我 一 个 忙,do me a favor)

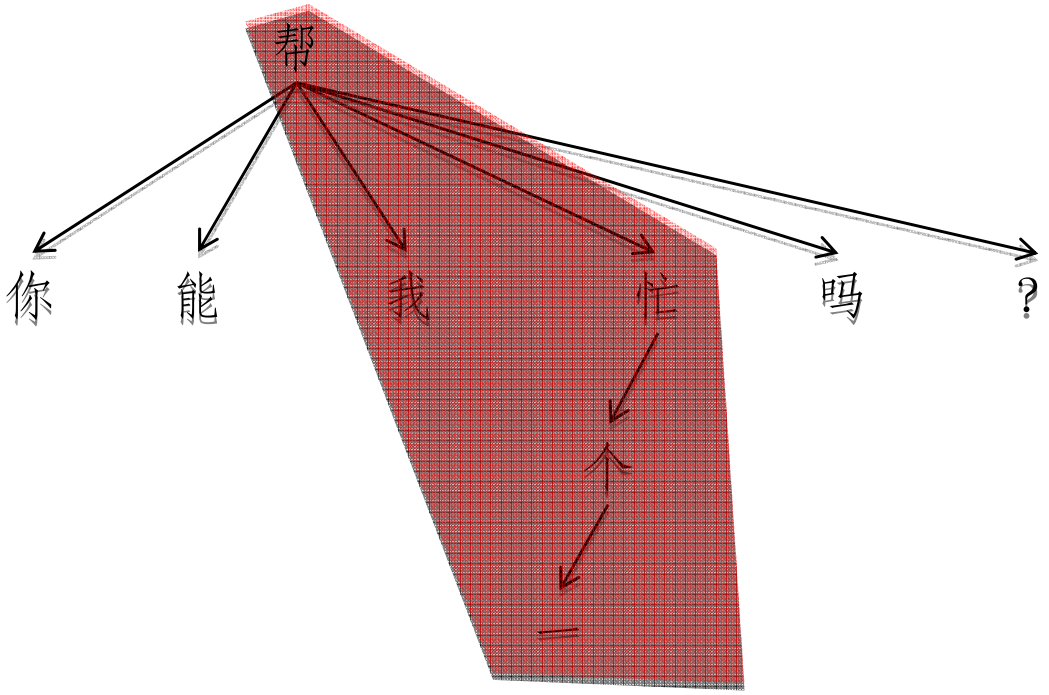


# 现有规则表示的不足
## ——短语兼容性（Dep2Str）

1.兼容短语
2.长距离调序

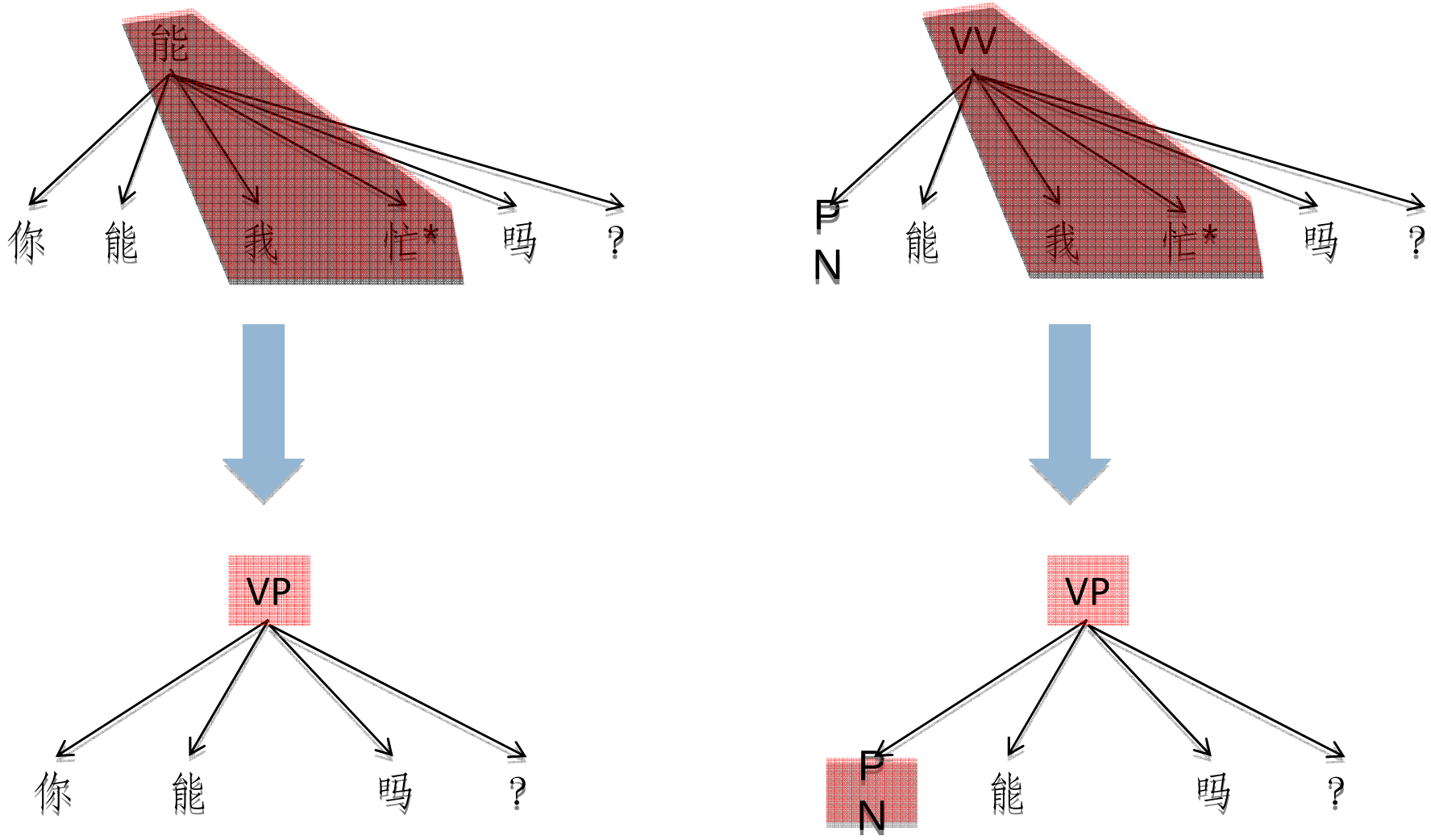



1.兼容短语
2.长距离调序




混合模型？
综合利用成分树和依存树包含的知识

# 现有规则表示的不足
## —短语兼容性（Dep2Str）



# 总结

- 规则表示决定了翻译模型的能力
- 规则表示需要同时处理调序、时态和语态
- 现有的规则表示还有进一步改进的空间（引入词义信息、词汇化上下文、结构化上下文）
- 成分树和依存树包含的知识具有一定的互补性
  - 理论上可以基于两种树结构设计新的规则表示从而更好地兼容两种树结构的有点

欢迎大家指导！

谢谢！