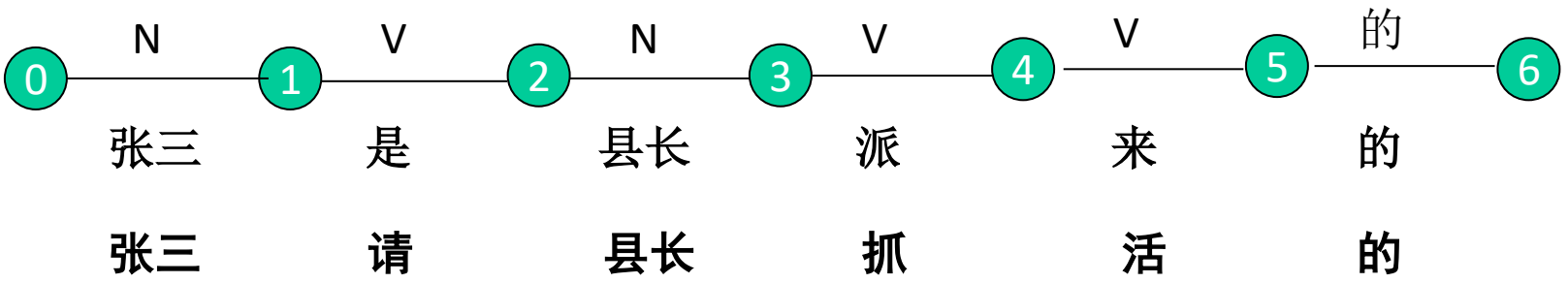
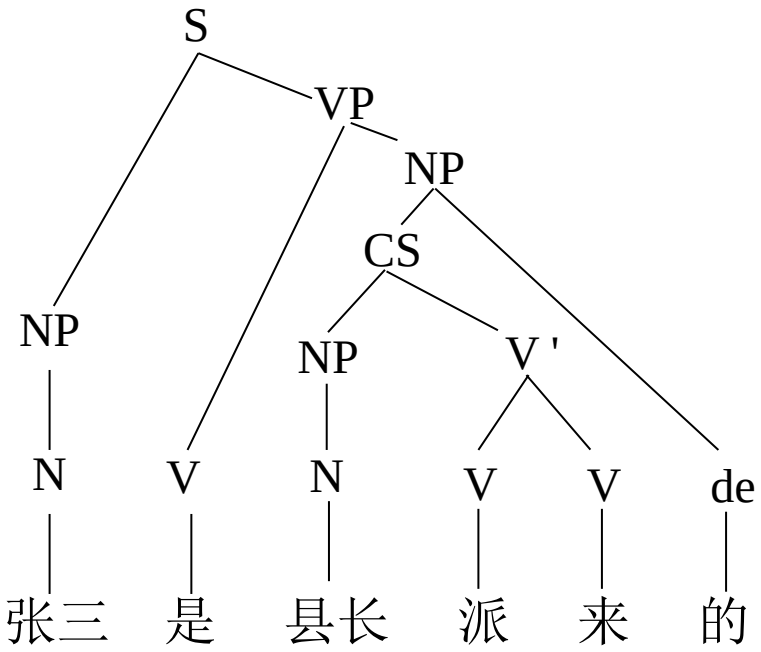
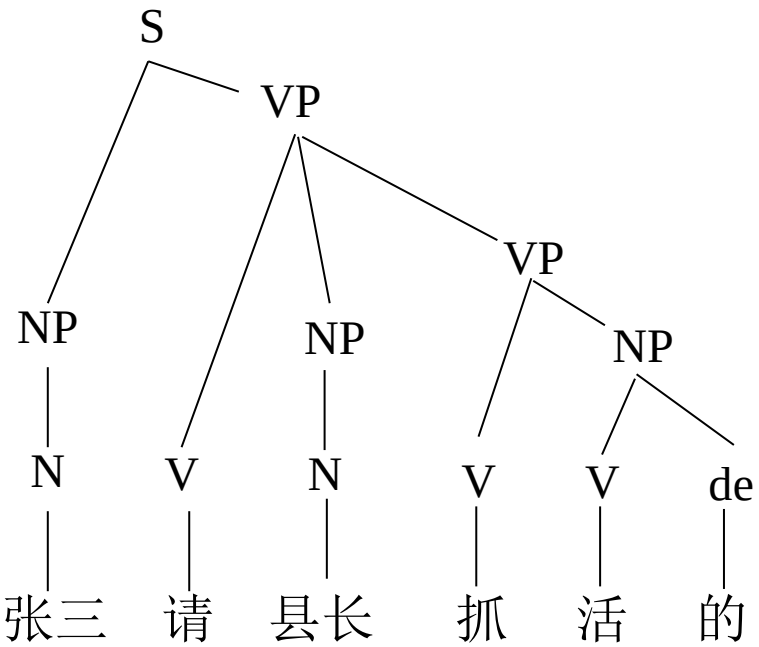
基于合一的Earley句法分析

詹卫东
zwd@pku.edu.cn

冯志伟、孙乐译《自然语言处理综论》电子工业出版社2005年版。第11章。
Daniel Jurafsky & James H. Martin, 2000, Speech and Language Processing, Pearson Education, Inc., Prentice Hall.


0
1
2
3
4
5
6
N
V
N
V
V
的
张三
是
县长
派
来
的
张三
请
县长
抓
活
的


S
VP
NP
CS
NP
V '
NP
N
V
N
V
V
de
张三
是
县长
派
来
的


S
VP
VP
NP
NP
NP
N
V
N
V
V
de
张三
请
县长
抓
活
的

对Earley算法进行改进

（1）根据重写规则所附的**合一约束**为一个节点生成**特征结构**

（2）对算法进程中的**状态**进行改进

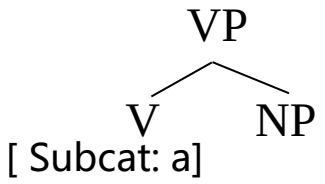
（3）对Predicator，Scanner，Completer等操作进行改进

（4）生成新状态时对**特征结构的蕴涵关系**进行检查

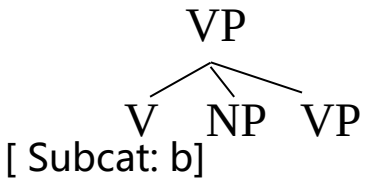
带合一约束的CFG规则

- 将重写规则中所附的约束转写为节点的特征结构

{R1} VP → V NP :: %V.Subcat = a



{R2} VP → V NP VP :: %V.Subcat = b



{R1}	{R2}	词语	词类	及物	兼语	Subcat
√	×	**是**	v	是	否	a
×	√	**请**	v	是	是	b
×	√	**派**	v	是	是	b
×	×	**来**	v	否	否	c

……

为Earley分析法中的状态增加特征结构字段

- 改进后的状态包含4部分：

  （1）重写规则，代表分析子树

  （2）子树分析的完成状况，用点标记 · 表示

  （3）子树完成部分与输入中词的位置对应关系

  （4）特征结构

$$\underbrace{S \rightarrow \bullet\, NP\ VP}_{(2)}\ ,\ \underbrace{[0,0]}_{(3)}\ ,\ \underbrace{DAG}_{(4)}$$

Predicator操作

对于状态 Z → α • X β [j, k] DAG_z 其中X是非终结符，对于语法中每条形如 X → γ DAG_x 的规则，都可以形成一个新状态： X → • γ [k, k] DAG_x

Scanner操作

对于状态 Z → α • X β [j, k] DAG_z 其中X是终结符，如果X与输入字符串中第k个字符匹配，就将词典中X的特征结构DAG_x跟DAG_z合一，

若成功，则形成一个新状态：Z → α X • β [j, k+1] $DAG_{X \cup Z}$

否则，不改变当前状态集。

Completer操作

对于一个已经“完成”的状态 Z → γ • [j, k] DAG_z，如果已有状态集合中有形如 X → α • Z β [i, j] DAG_x这样的状态，就将DAG_z跟DAG_x进行合一运算，

若成功，则形成一个新状态：X → α Z • β [i, k] $DAG_{X \cup Z}$

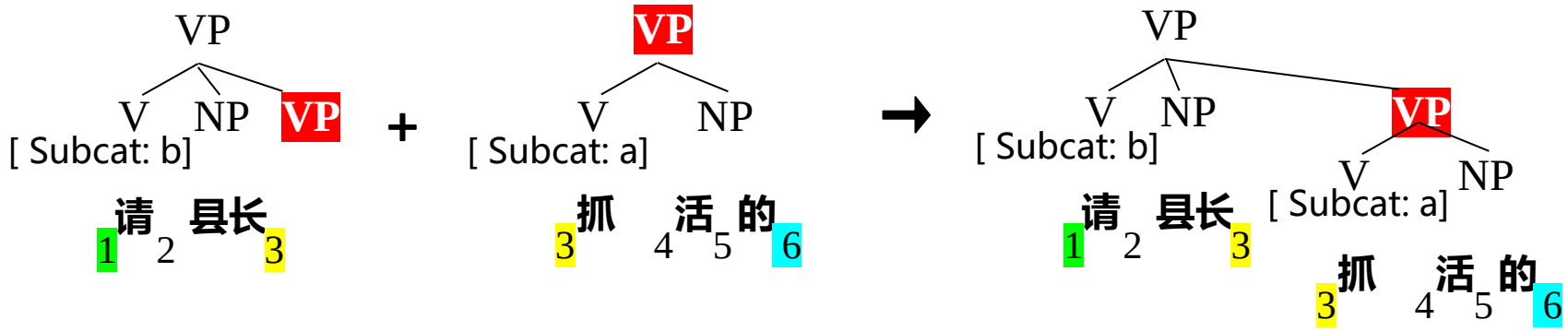
否则，不改变当前状态集。

Completer示例

0 张三 1 请 2 县长 3 抓 4 活 5 的 6

VP → V NP • VP [1,3] %V.subcat = b

VP → V NP • [3,6] %V.subcat = a



产生新状态时检查DAG蕴涵关系

- 在Earley算法中，如果新产生的状态和状态集合中的某个状态**相同**，则新状态**不加入状态集**。
- 在基于合一的Earley算法中，如果新状态的特征结构被状态集中已有状态的特征结构**蕴涵**，则新状态**不加入状态集**。

VP → V•NP :: %V.subcat=a　　预测1：NP（无要求）

VP → V•NP VP :: %V.subcat=b
%NP.语义类=人　　~~预测2：NP（语义类=人）~~

基于合一的Earley分析算法

设输入字符串长度为 n, 字符间隔可记做0，1，2，…，n

1　将文法规则中形如 S→ α DAG_s 的规则形成为**种子状态**：

　　<S→ · α [0, 0] DAG_s> 加入到状态集合中；

2　对当前分析句子的**每个词**，**循环**执行：

　对状态集中的**每个状态**，**循环**执行：

　2.1 如果当前状态是[**未完成状态**]，且点后是**非终结符**，则执行Predicator；

　2.2 如果当前状态是[**未完成状态**]，且点后是**终结符**，则执行Scanner；

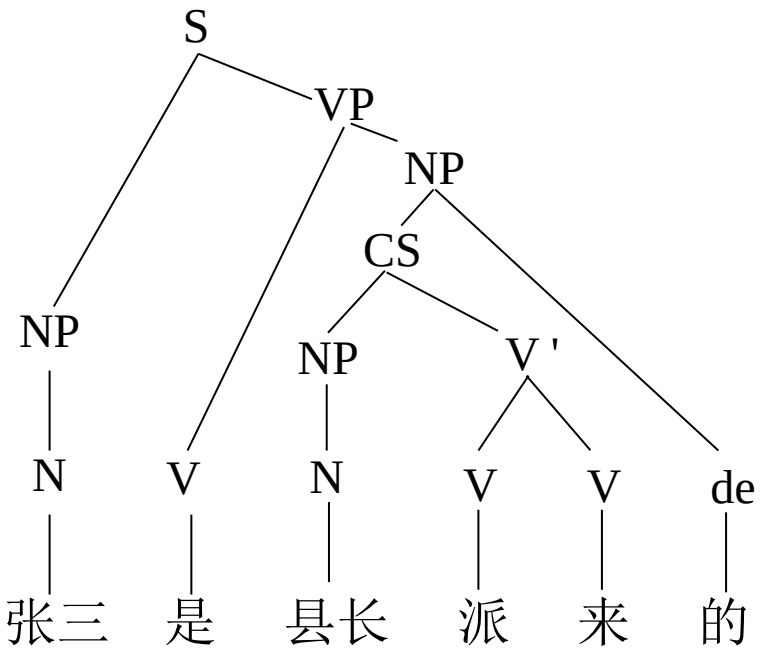
　2.3 如果当前状态是[**完成状态**]，则执行Completer；

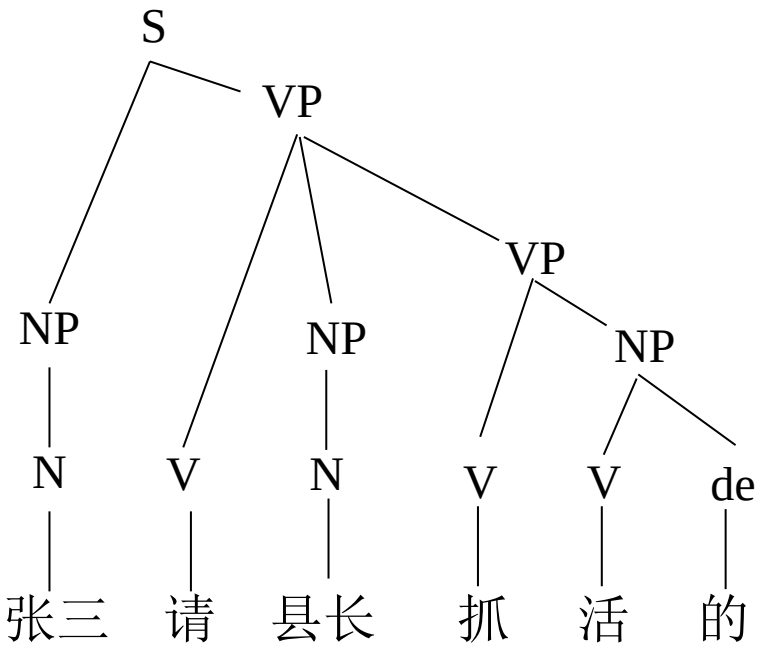
3　如果最后得到形如<S→ α · [0, n] DAG_s>这样的状态，那么输入字符串被接受为**合法**的句子，否则句子不合法。

基于合一的Earley分析法示例

(1) S → NP VP
(2) NP → N
(3) NP → CS 的
(4) CS → NP V '
(5) VP → V NP
(6) V ' → V V

(7) NP → V 的
(8) VP → V NP VP



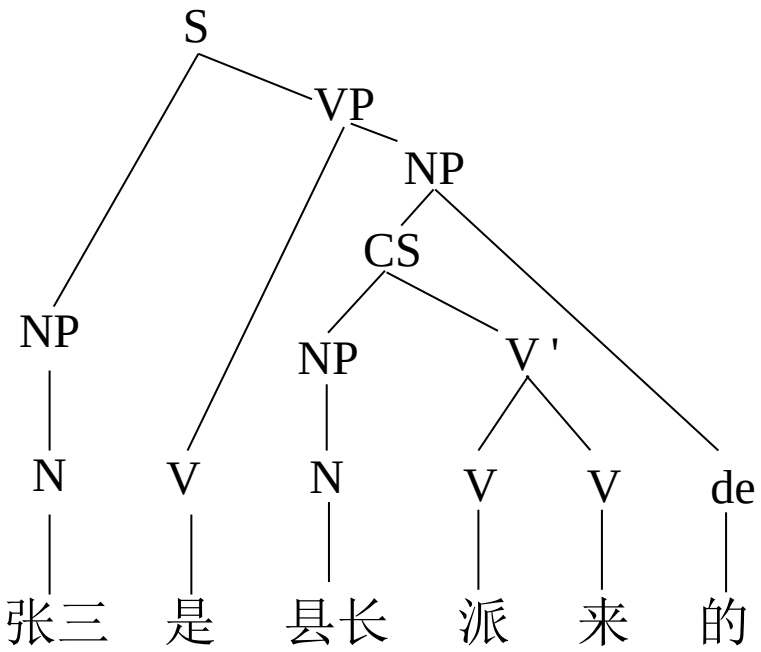


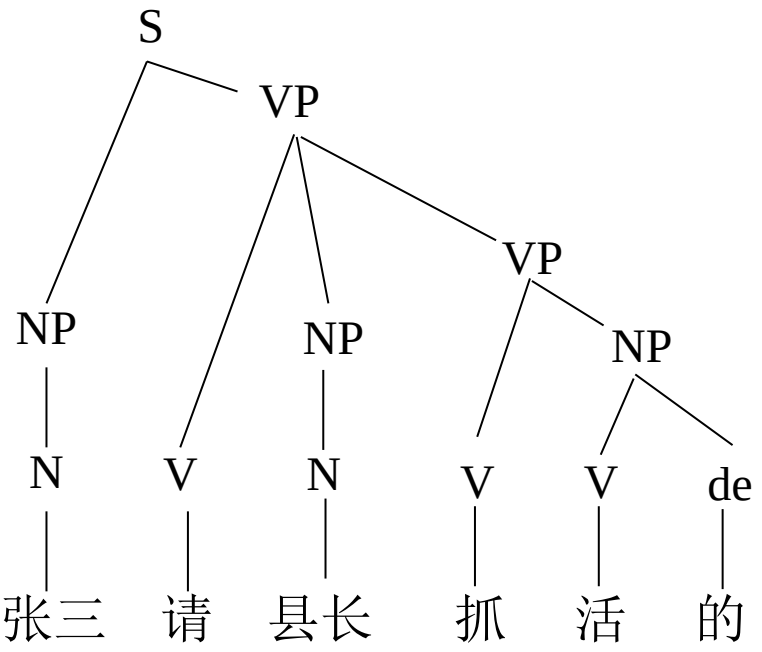
基于合一的Earley分析法示例

① 是 V [subcat:a]
② 派 V [subcat:b]
③ 来 V [subcat:c]
④ 请 V [subcat:b]
⑤ 抓 V [subcat:a]
⑥ 活 V [subcat:d]

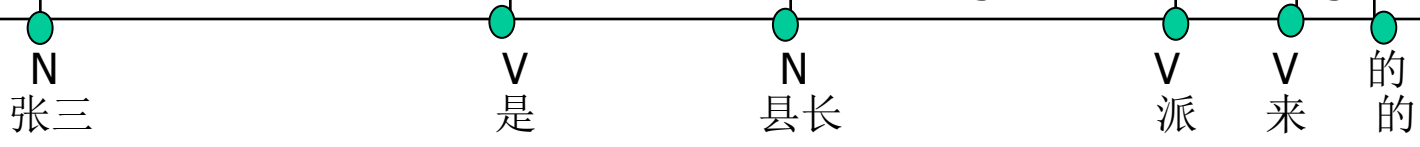
(1) S → NP VP
(2) NP → N
(3) NP → CS 的
(4) CS → NP V '
(5) VP → V NP :: %V.subcat=a
(6) V ' → V V :: %%V.subcat=c

(7) NP → V 的
(8) VP → V NP VP :: %V.subcat=b
%NP.语义类=人

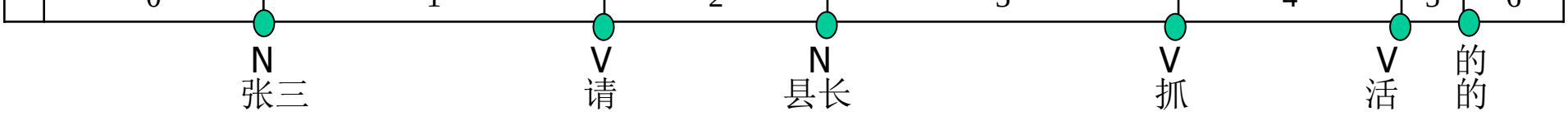




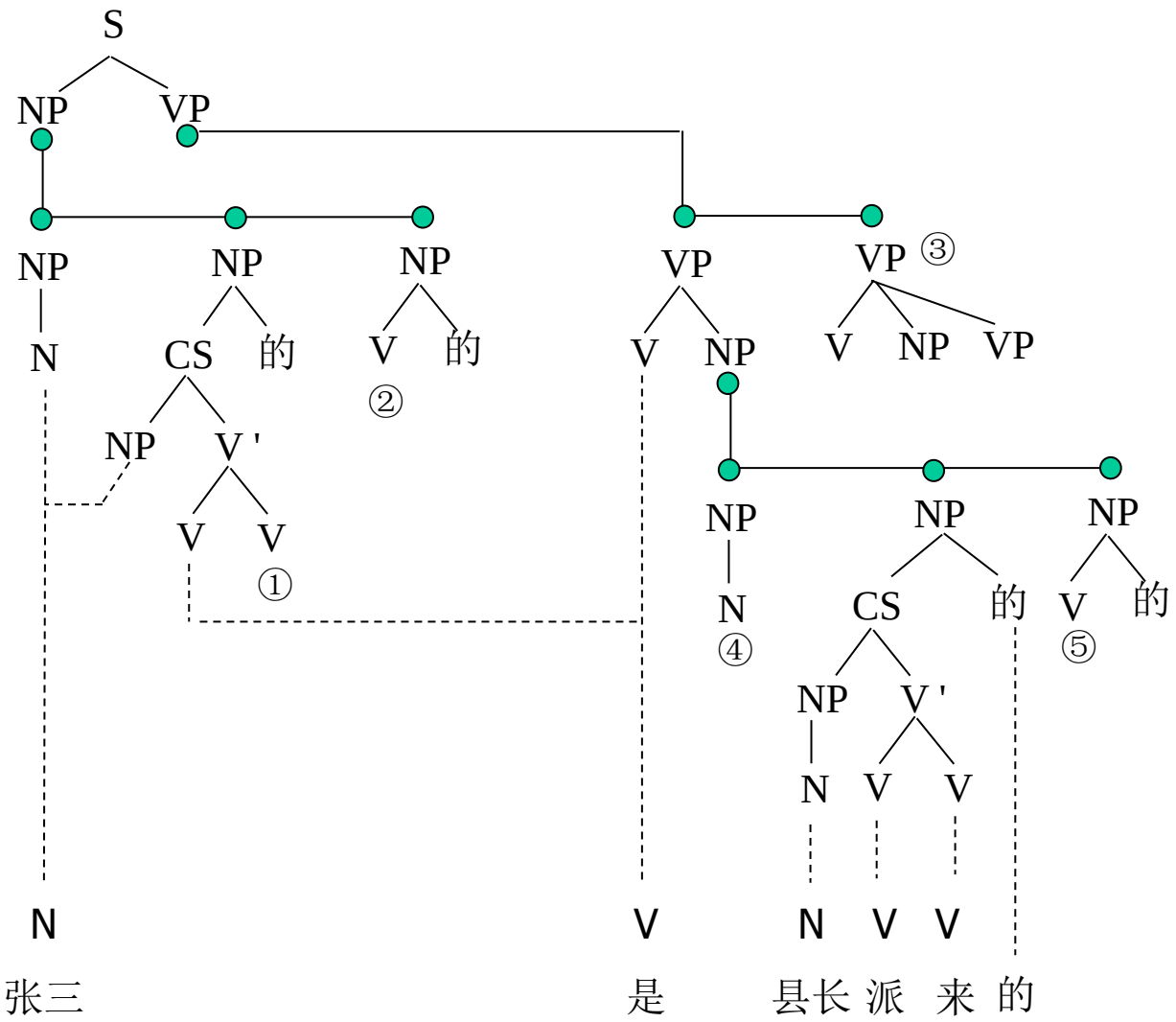
6	S → NP VP ·	VP → V NP · $dag_{①-(5)}$	NP → CS 的 ·				扩展 扫描
5			NP → CS · 的 CS → NP V ' ·	V ' → V V · $dag_{②-(6)-③}$			扩展 扫描
4				V ' → V · V $dag_{②-(6)}$			扫描
3	S → NP VP ·	VP → V NP · $dag_{①-(5)}$	CS → NP · V ' NP → N ·	V ' → · V V $dag_{(6)}$			预测 扩展 扫描
2		~~VP → V· NP VP $dag_{①-(8)}$~~ VP → V· NP $dag_{①-(5)}$ V ' → V· V $dag_{①-(6)}$	CS → · NP V ' NP → · V 的 NP → · N NP → · CS 的				预测 扫描
1	S → NP · VP CS → NP · V ' NP → N ·	VP → · V NP VP $dag_{(8)}$ VP → · V NP $dag_{(5)}$ V ' → · V V $dag_{(6)}$					预测 扩展 扫描
0	CS → · NP V ' NP → · V 的 NP → · N NP → · CS 的 S → · NP VP						预测 种子
	0	1	2	3	4	5	6



6	S → NP VP ·	VP →V NP VP · dag④-(8)		VP → V NP · dag⑤-(5) VP → V NP · VP dag⑤-(8)	CS→ NP · V ' NP → V 的 ·		扩展 扫描
5				~~V ' → V V · dag⑤-(6)-⑥~~	NP → V · 的		扫描
4				V ' → V · V dag⑤-(6) VP → V · NP dag⑤-(5) VP → V · NP VP dag⑤-(8)	CS → · NP V ' NP → · V 的 NP → · N NP → · CS 的		预测 扫描
3		VP →V NP · VP dag④-(8)	CS → NP · V ' NP → N ·	V ' → · V V dag(6) VP → · V NP dag(5) VP → · V NP VP dag(8)			预测 扩展 扫描
2		VP → V · NP VP dag④-(8) ~~VP → V · NP dag④-(5)~~ V ' → V · V dag④-(6)	CS → · NP V ' NP → · V 的 NP → · N NP → · CS 的				预测 扫描
1	S → NP · VP CS → NP · V ' NP → N ·	VP → · V NP VP dag(8) VP → · V NP dag(5) V ' → · V V dag(6)					预测 扩展 扫描
0	CS → · NP V ' NP → · V 的 NP → · N NP → · CS 的 S → · NP VP						预测 种子
	0	1	2	3	4	5	6



Earley算法构造分析树示意图



① 因为“是”的后面不是V，该节点“到此为止”。

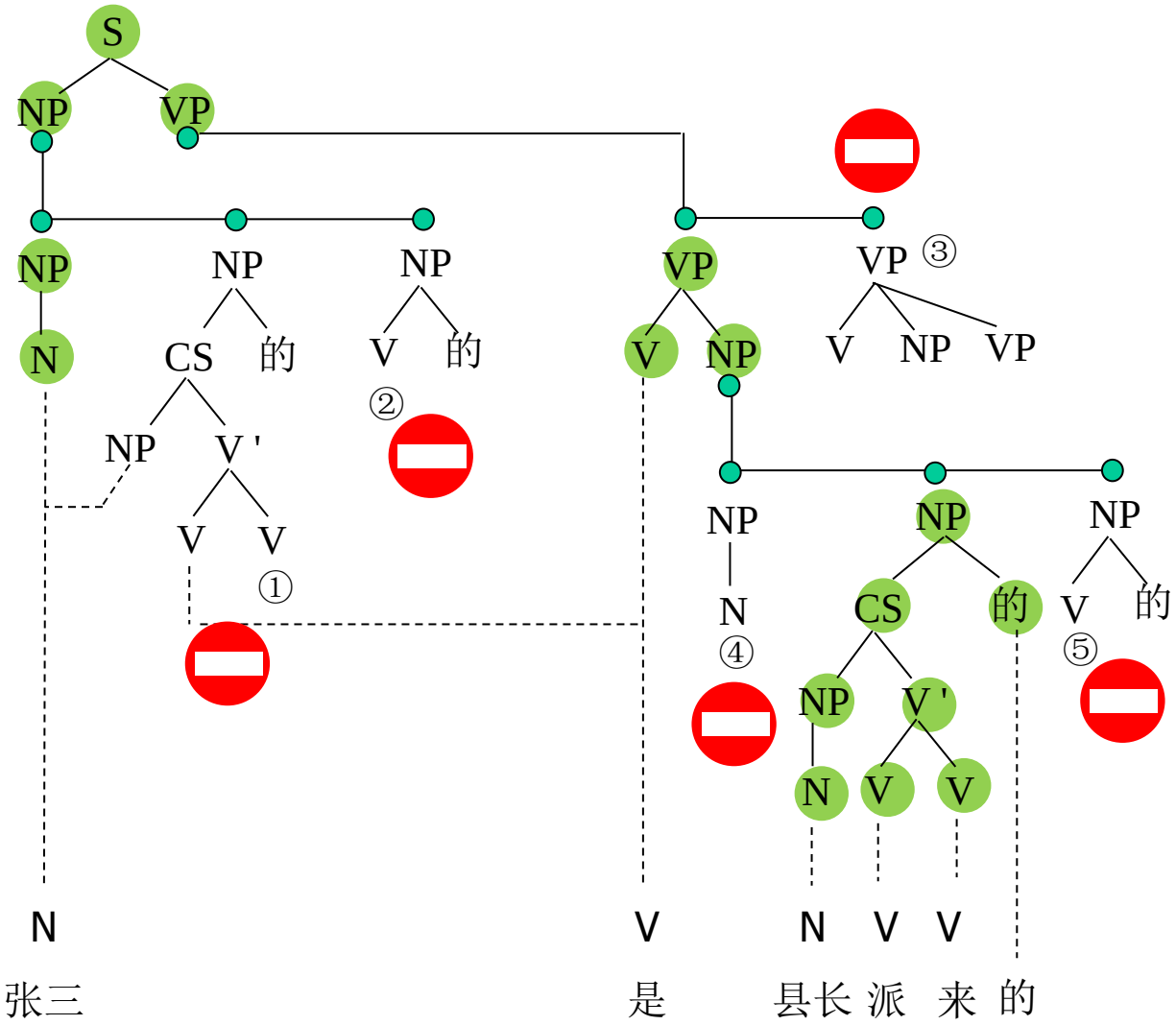
② 起始位置不是“V”。

③该规则的约束要求V的子类为b，而“是”的子类特征值为a，合一失败。

④ N如果跟“县长”匹配成功，则分析完毕，但句子并未结束。

⑤ V 跟“县长/N”不匹配

Earley算法构造分析树示意图



① 因为“是”的后面不是V，该节点“到此为止”。

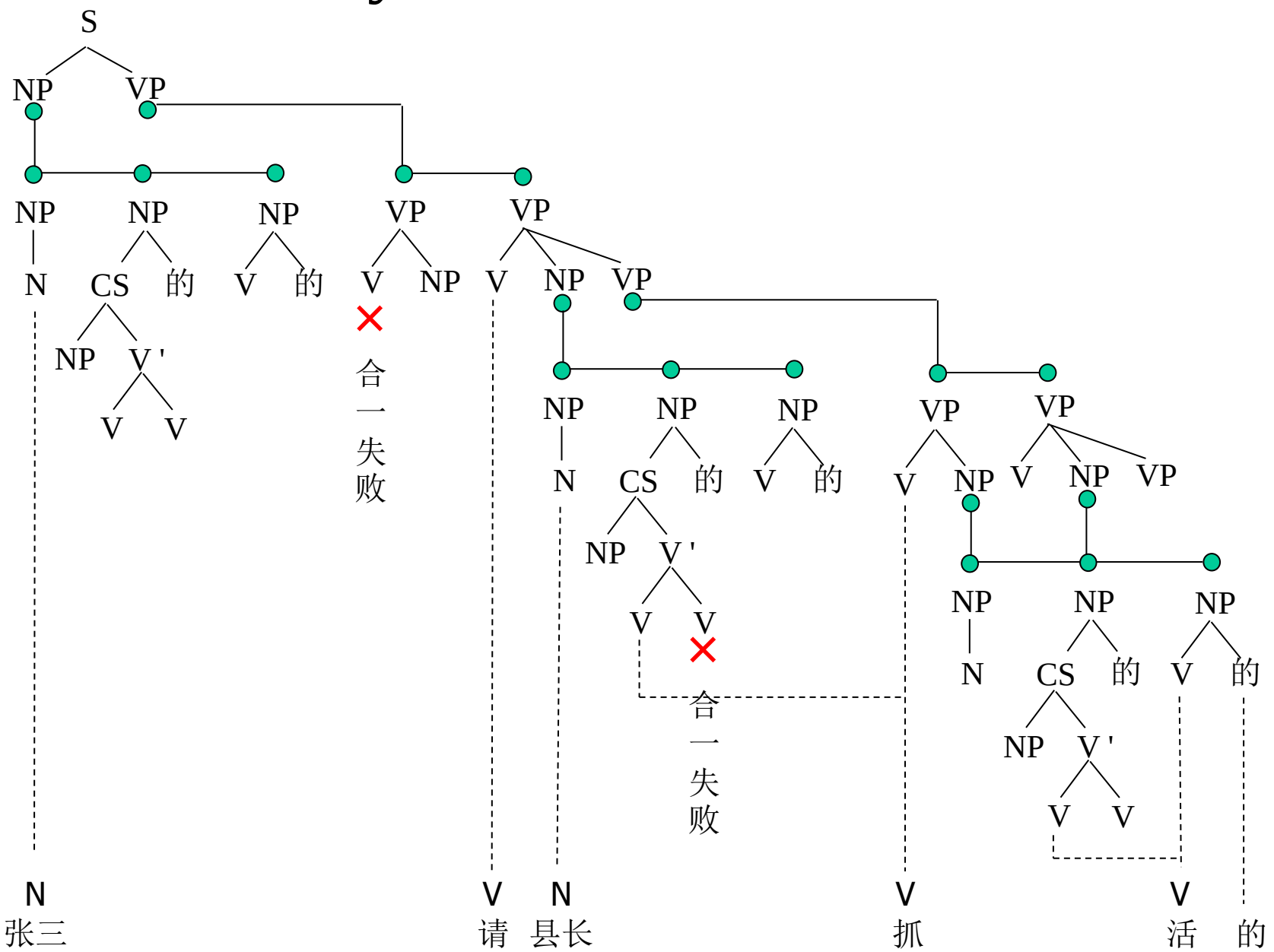
② 起始位置不是“V”。

③该规则的约束要求V的子类为b，而“是”的子类特征值为a，合一失败。

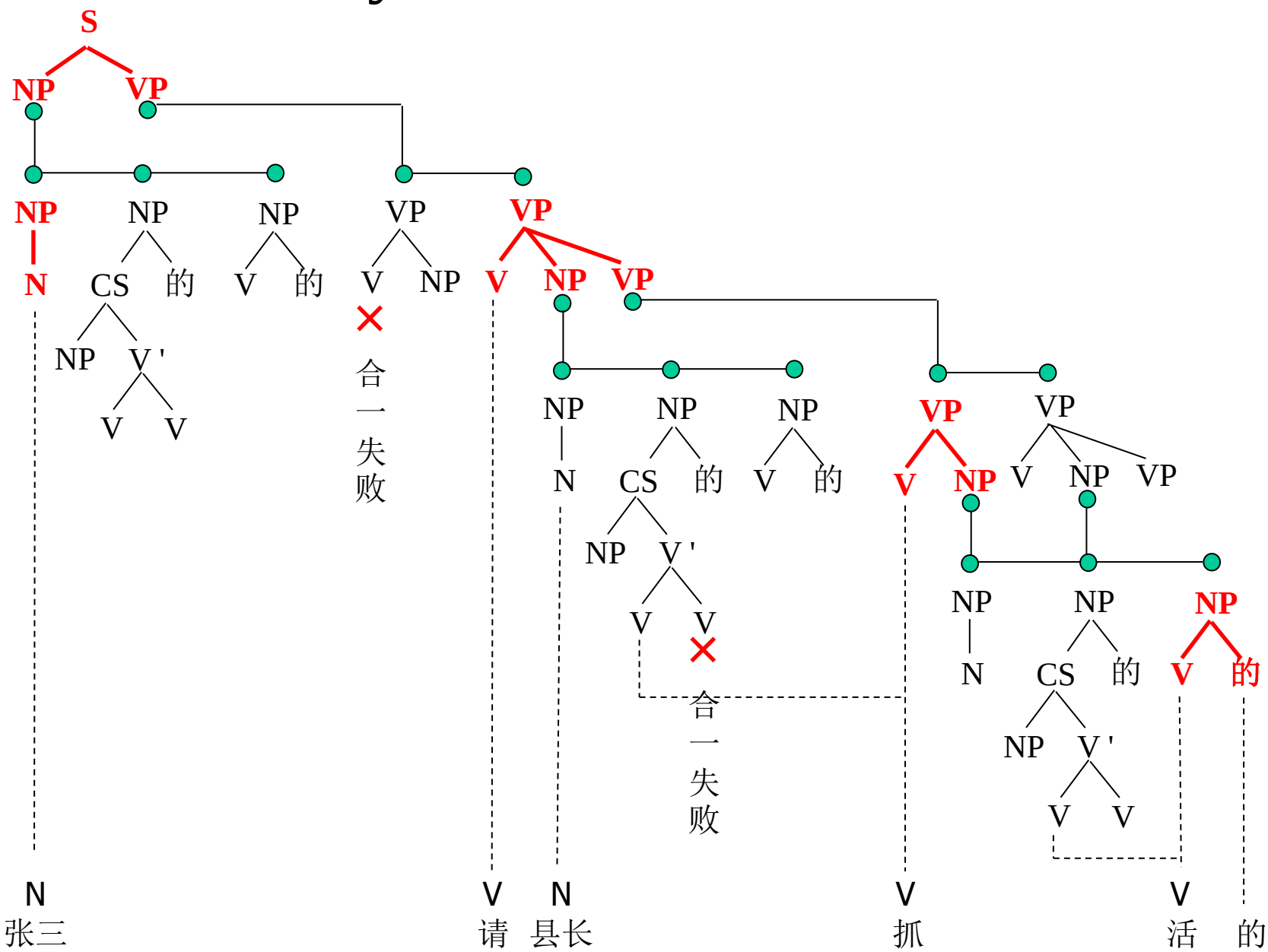
④ N如果跟“县长”匹配成功，则分析完毕，但句子并未结束。

⑤ V 跟“县长/N”不匹配

Earley算法构造分析树示意图



Earley算法构造分析树示意图



练 习

请构造带有合一描述的上下文无关文法规则集和词典，然后用Earley算法分析下面句子的结构：

1. 张三 请 李四 抓 活 的 回来
2. 张三 说 李四 是 要饭 的 装阔