计算语言学

第14讲 句法分析II

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

内容提要

- 上下文无关语法的分析算法
  - 富田算法（Tomita算法）
  - 线图分析算法（Chart算法）
- 概率上下文无关语法
- 组块分析与部分分析

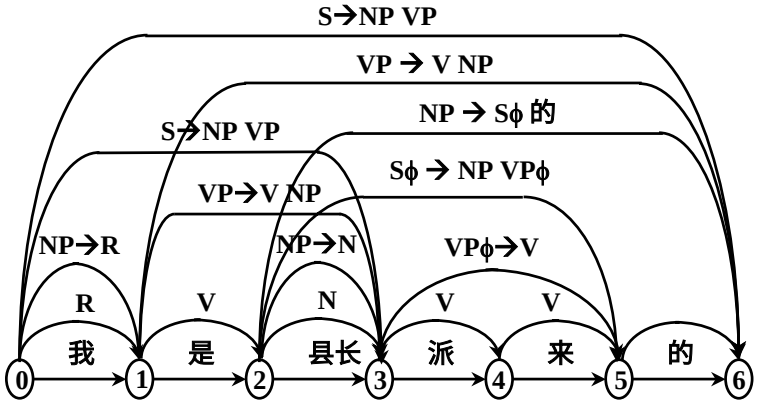
线图（Chart）分析算法

- 线图分析算法 Chart Parsing Algorithm
- 线图分析法的核心是线图（Chart）表示法，线图表示法具有简单、直观的特点；
- 通过修改线图分析法的分析策略，可以方便地模拟很多种分析算法，如自顶向下的分析方法、自底向上的分析方法、左角分析方法等等。

线图表示法 1

- 线图是一个无环有向图（DAG），其中：
  - **结点**：输入句子中词与词之间的每一个间隔为一个结点；结点的标记往往用一个序号来表示；
  - **边（弧）**：对应于句子中的一个短语，边两端的结点给定了短语的边界，边的方向总是从左到右。边上面不仅要标记短语的类型，还需要标记产生该短语的规则。
  - **说明**：在汉语分析中，为了兼容词语切分的歧义，常常将汉字之间的间隔作为一个结点

线图表示法 2



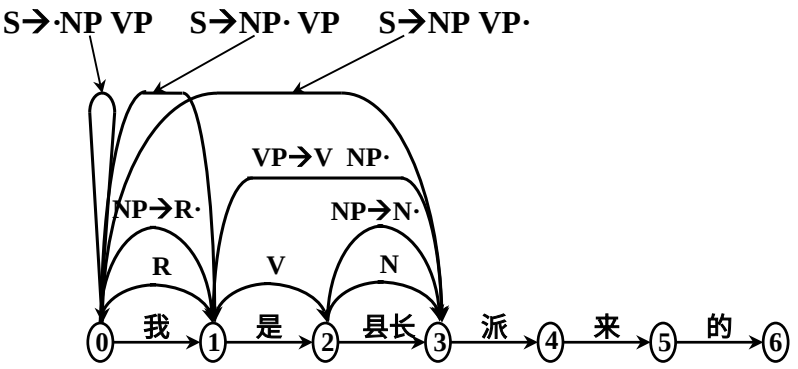
活跃边与非活跃边 1

- 上述记录分析成功的短语的边称为非活跃边。
- 在线图中，还有另一种形式的边，用于记录一条规则不完全分析的结果，称为活跃边，如下所示：

记录方式	边状态	匹配程度	起点	终点	对应词串
<0,0, S→·NP VP>	活跃	**S → ·NP VP**	0	0	
<0,1, S→NP· VP>	活跃	**S → NP· VP**	0	1	我
<0,3, S→NP VP·>	非活跃	**S → NP VP·**	0	3	我是县长

- 活跃边的引入，可以减少规则匹配中的冗余操作，提高句法分析的效率。

活跃边与非活跃边 2



日程表(Agenda)

- 在线图分析算法中，除了“线图(Chart)”以外还有一个重要的数据结构，称为“日程表(Agenda)”
- Chart分析的过程就是一个不断产生新的边的过程。但是每一条新产生的边并不能立即加入到Chart中，而是要放到日程表(Agenda)中
- 日程表(Agenda)实际上是一个边的集合，用于存放已经产生，但是还没有加入到Chart中的边。
- 日程表(Agenda)中边的排序和存取方式，是Chart算法执行策略的一个重要方面

线图分析算法的基本流程

Chart算法就是一个由日程表驱动的不断循环的过程：

1. 按照**初始化策略**初始化Agenda
2. 如果Agenda为空，那么分析失败
3. 每次按照**日程表组织策略**从Agenda中取出一条边
4. 如果取出的边是一条非活跃边，而且覆盖整个句子，那么返回成功
5. 将取出的边加入到Chart中，执行**规则匹配策略**和**规则调用策略**，将产生的新边又加入到Agenda中
6. 返回第(2)步

线图分析算法：初始化策略

- Chart分析算法开始执行以前，要先将Agenda初始化
- 对于不同的句法分析策略，初始化策略也不相同

自底向上分析的规则调用策略

– 将所有单词（含词性）边加入到Agenda中。

自顶向下分析的规则调用策略

– 将所有单词（含词性）边加入到Agenda中；
– 对于所有形式为S→W的规则，产生一条形式为 <0, 0, S→·W> 的边，并加入到Agenda中；

线图分析算法：规则匹配策略

- 在Chart算法中，边是逐条从Agenda中加入到Chart中的
- 将每一条边从Agenda中取出并加入到Chart中时，都要执行以下**规则匹配策略**：
  - 如果新加入一条活跃边形式为：<i, j, A→W1· B W2>，那么对于Chart中所有形式为：<j, k, B→W3>的非活跃边，生成一条形式为<i, j, A→W1 B· W2>的新边，并加入到Agenda中；
  - 如果新加入一条非活跃边形式为：<j, k, B→W3>，那么对于Chart中所有形式为：<i, j, A→W1· B W2>的活跃边，生成一条形式为<i, j, A→W1 B· W2>的新边，并加入到Agenda

  上面A、B为非终结符，W1、W2、W3为终结符和非终结符组成的串，其中W1、W2允许为空，W3不允许为空

线图分析算法：规则调用策略

- 对于不同的句法分析策略，规则调用策略也不相同：

  自底向上分析的规则调用策略
  - 如果要加入一条形式为<i, j, C→W1·>的边到Chart中，那么对于所有形式为B→C W2的规则，产生一条形式为<i, i, B→· C W2>的边加入到Agenda中

  自顶向下分析的规则调用策略
  - 如果要加入一条形式为<i, j, C→W1·B W2>的边到Chart中，那么对于所有形式为B→W的规则，产生一条形式为<j, j, B→·W>的边，并加入到Agenda中

线图分析算法：日程表组织策略

- 通过日程表组织的不同策略，可以分别实现深度优先和广度优先等搜索方法

  深度优先的日程表组织策略

  – 将日程表按照堆栈的形式，每次从日程表中取出最后加入的结点；

  广度优先的日程表组织策略

  – 将日程表按照队列的形式，每次从日程表中取出最早加入的结点；

线图分析算法：细节处理

- 前面的讨论中忽略了两个细节，在实现一个系统时应该考虑到：
  – 考虑到可能通过多种途径生成一条完全相同的边，所以每次从Agenda中取出一条新边加入Chart时，要先检查一下Chart中是否已经有相同的边，如果有，那么删除这条边，直接进入下一个循环
  – 为了生成最后的句法结构树，每一条边中还应该记录该条边的子句法成分所对应的边

线图分析算法：总结

- Chart算法具有直观和灵活的特点；
- 通过修改分析过程中的一些具体策略，Chart分析算法可以模拟很多种其他句法分析算法。
- 白硕和张浩在“角色反演算法”（软件学报，已录用）中，把Tomita算法中“向前看（look ahead）”的思想结合到Chart分析算法中，提出了一种“角色反演算法”，可以减少Chart分析算法中垃圾边的数量而又不影响最后的分析结果，提高分析的效率。

复习思考题

- 尝试修改Chart算法中的分析策略，以实现新的句法分析算法
- 尝试修改Chart算法中活跃边的表示方法，实现从右到左的句法分析或者从中心词向两边扩展的句法分析