自然语言处理

Natural Language Processing

权小军 教授
中山大学数据科学与计算机学院
quanxj3@mail.sysu.edu.cn

课程回顾

概述

- 句法分析是自然语言处理中的基础性工作，它分析句子的句法结构（主谓宾结构）和词汇间的依存关系（并列，从属等）；

概述

- 句法分析是自然语言处理中的基础性工作，它分析**句子的句法结构**（主谓宾结构）和**词汇间的依存关系**（并列，从属等）；
- 句法分析可以为语义分析、情感倾向、观点抽取等NLP应用场景打下坚实的基础。

句法分析是语言理解的重要基础

深度学习时代也不例外！

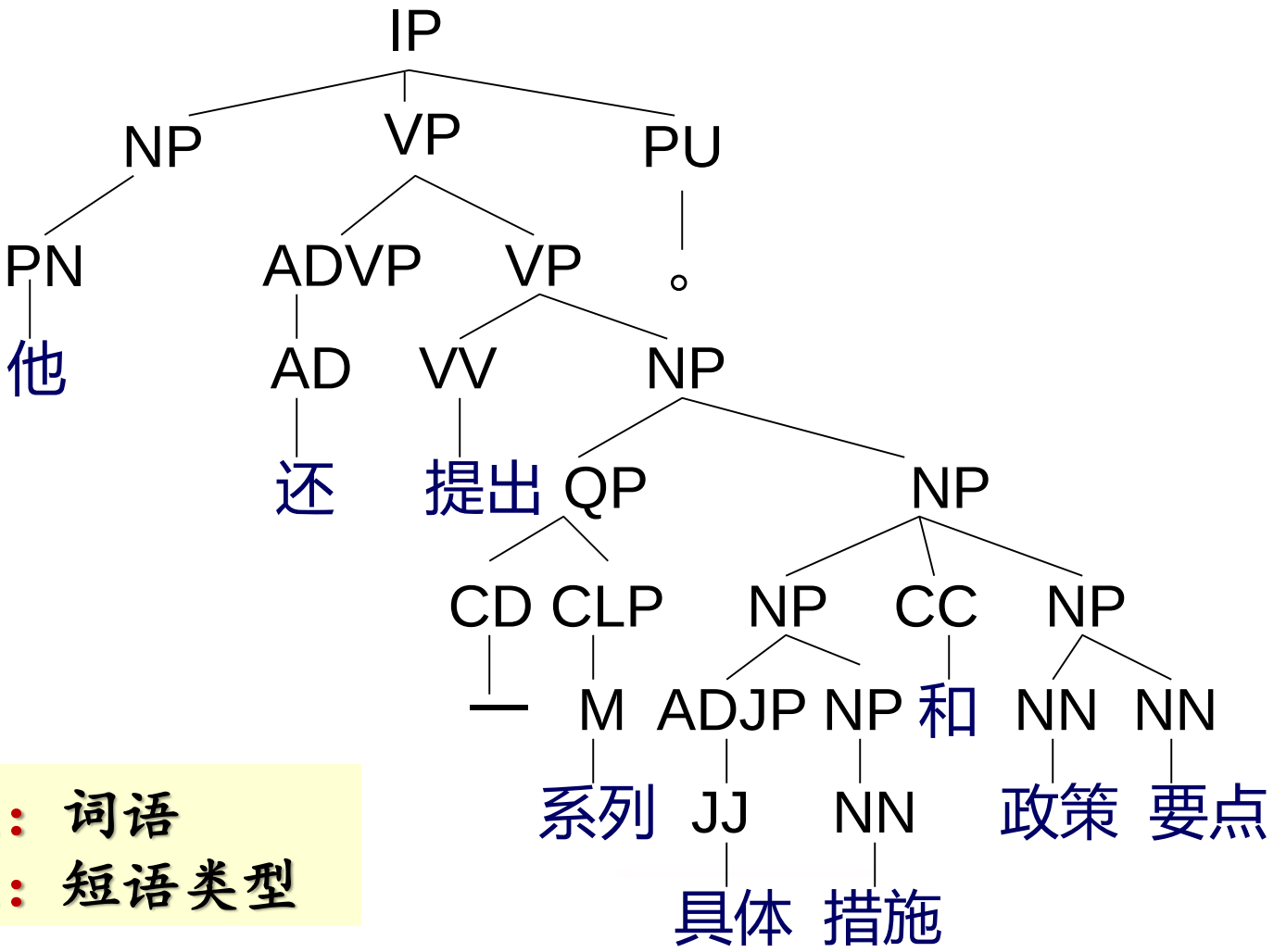
概述

1、您转的这篇文章很无知

2、您转这篇文章很无知

短语结构分析

树状表示：



- 叶子节点：词语
- 内部节点：短语类型

短语结构分析

符号解释：

NP：名词短语
VP：动词短语
PU：断句符，通常是句号、问号、感叹号等标点符号
PP：介词短语
CP：由‘的’构成的表示修饰性关系的短语
ADVP：副词短语
ADJP：形容词短语
DP：限定词短语
QP：量词短语
NN：常用名词
NT：时间名词
PN：代词
VV：动词
……

短语结构分析

- 目标：实现高正确率、高鲁棒性(robustness)、高速度的自动句法分析过程；
- 困难：自然语言中存在大量的复杂的结构歧义(structural ambiguity)；

短语结构分析

- 基本方法和开源的句法分析器：
  - 基于CFG规则的分析方法
    - CFG: Context-Free Grammar (上下文无关文法)
    - 代表：线图分析法(chart parsing)

延伸

  - 基于 PCFG 的分析方法
    - PCFG: Probabilistic Context-Free Grammar (概率上下文无关文法)

1、概述

2、短语结构分析

a) 上下文无关文法

上下文无关文法（CFG）

- **CFG由一系列规则组成，每条规则给出了语言中的某些符号可以被组织或排列在一起的方式.**

符号被分成两类：

- 终结点(叶子节点)：就是指单词，例如 book；
- 非终结点(内部节点)：句法标签，例如 NP 或者 NN；

规则是由一个“→”连接的表达式：

- 左侧：只有一个 non-terminal；
- 右侧：是一个由符号组成的序列；

上下文无关文法（CFG）

CFG示例:

- 符号：
  - 终结点: rat, the, ate, cheese;
  - 非终结点: S, NP, VP, DT, VBD, NN;
- 规则：

S → NP VP
NP → DT NN
VP → VBD NP
DT → the
NN → rat
NN → cheese
VBD → ate

基于上下文无关文法的句法分析

基于上下文无关文法的句法分析基于预定义的语法，为语句生成恰当的句法树，要求该树：

- ✓ 符合给定语法；
- ✓ 叶子节点包含所有的词；

符合这样条件的树通常有很多！

1、概述

2、短语结构分析

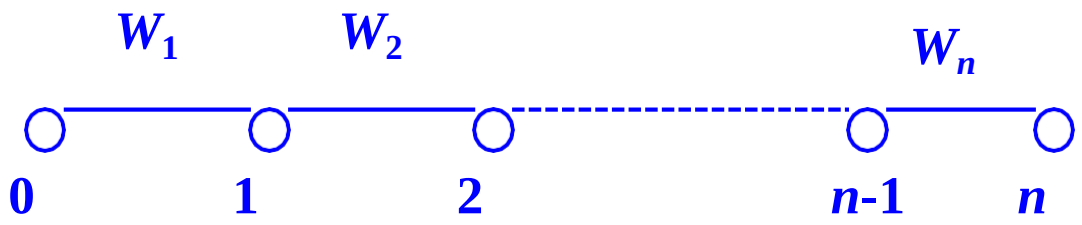
a) 上下文无关文法

b) 线图分析法

线图分析法

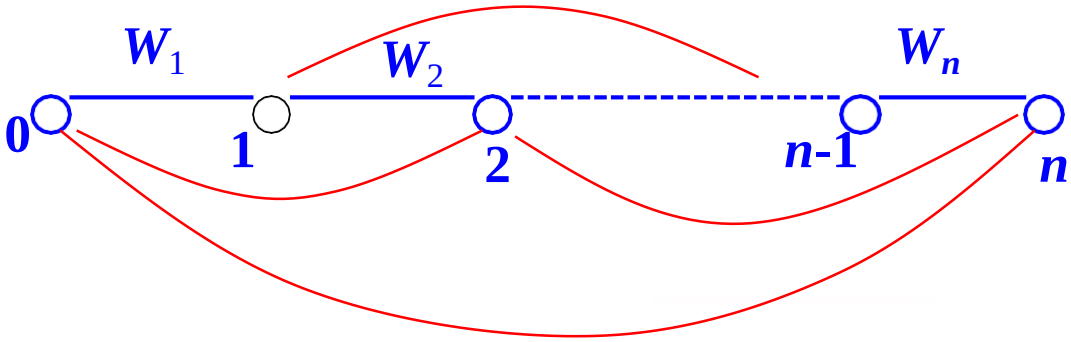
❑ 自底向上的线图分析算法

- 给定一组 CFG 规则： $XP \rightarrow \alpha_1 \ldots \alpha_n \quad (n \geq 1)$
- 给定一个句子的**词性序列**：$S = W_1 W_2 \cdots W_n$
- 构造一个线图：一组结点和边的集合；

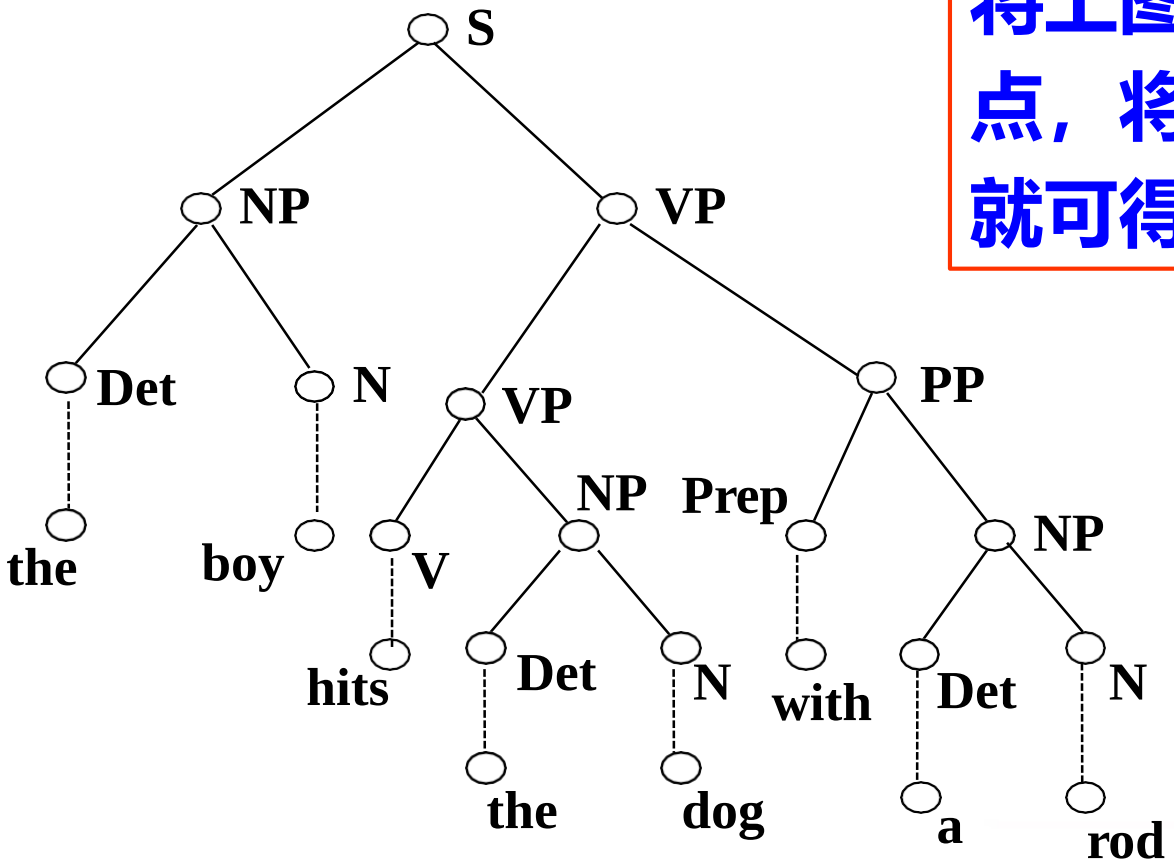


线图分析法

执行：查看任意相邻几条边上的词性串是否与某条规则的右部相同，如果相同，则增加一条新的边跨越原来相应的边，新增加边上的标记为这条规则的头(左部)。重复这个过程，直到没有新的边产生。



线图分析法



将上图中的边改为结点，将结点改为边，就可得到一棵句法树

1、概述

2、短语结构分析

a) 上下文无关文法

b) 线图分析法

c) 概率上下文无关文法

概率上下文无关文法

- 根据一个CFG构建语法分析树，往往不止一个;
- 对于可能产生多种语法分析结果的问题，怎么办?
- 引入概率上下文无关文法（PCFG， Probabilistic context-free grammar）：**给每棵树计算一个概率**!

概率上下文无关文法

❑ PCFG 规则

形式：$A \rightarrow \alpha \quad [p]$

- NP → DT NN [p = 0.45]
- NN → leprechaun [p = 0.0001]

概率上下文无关文法

❑ 例-1：S → NP VP, [1.00]　NP → NP PP, [0.40]

NP → astronomers, [0.10]

NP → ears, [0.18]　NP → saw, [0.04]

NP → stars, [0.18]　NP → telescopes, [0.10]

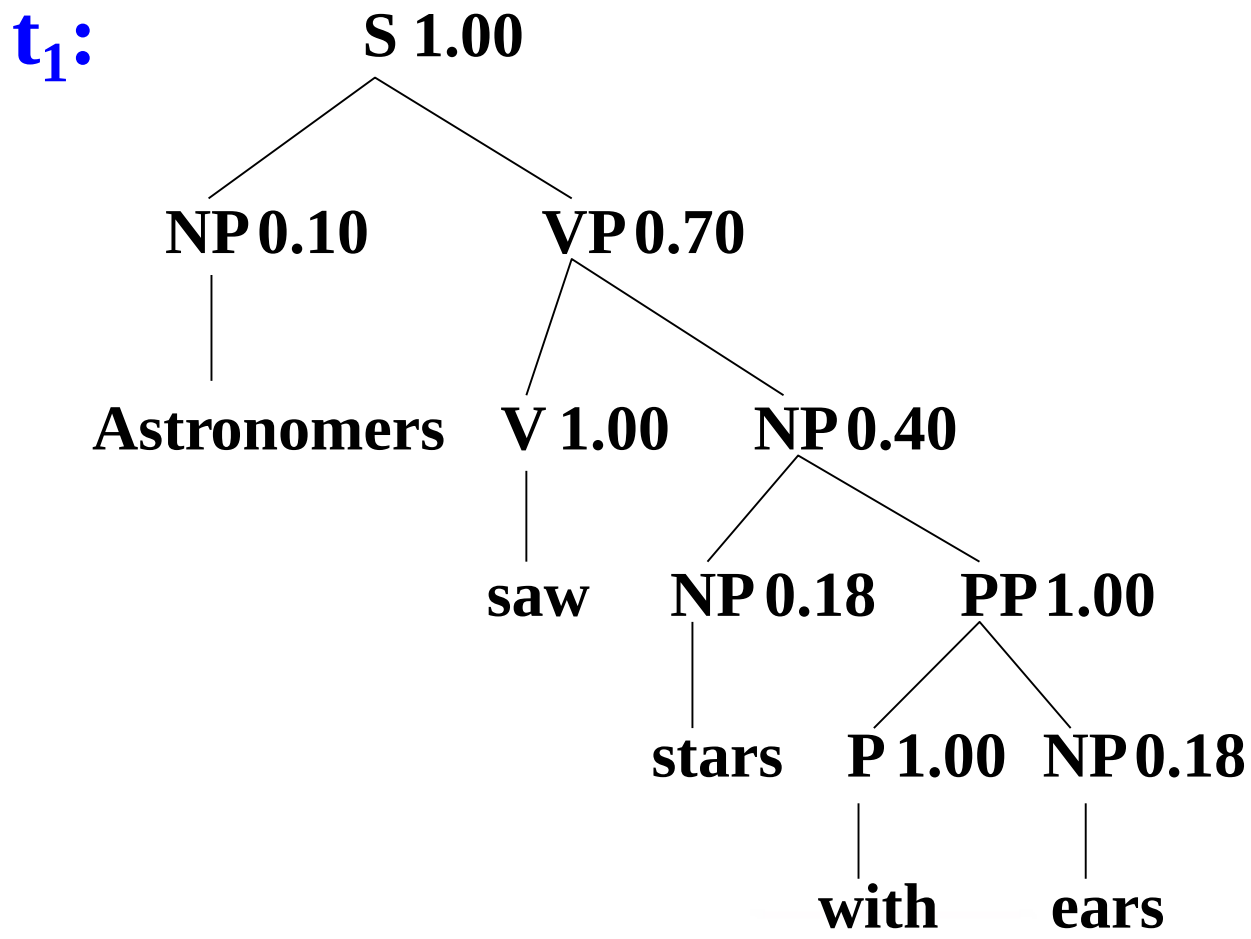
PP → P NP, [1.00]　P → with, [1.00]

VP → V NP, [0.70]　VP → VP PP, [0.30]

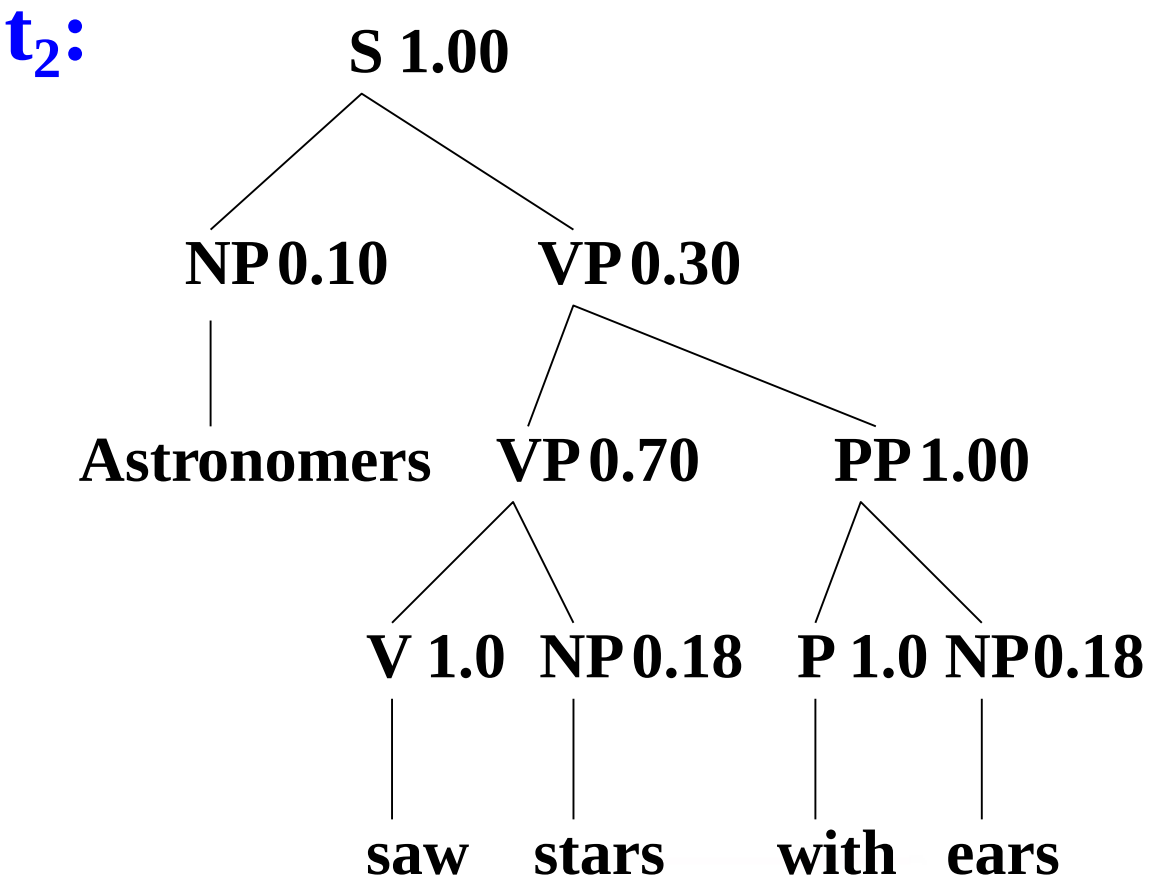
V → saw, [1.00]

给定句子 S：***Astronomers saw stars with ears.***

概率上下文无关文法


t_1:
S 1.00
NP 0.10
VP 0.70
Astronomers
V 1.00
NP 0.40
saw
NP 0.18
PP 1.00
stars
P 1.00
NP 0.18
with
ears

概率上下文无关文法


t_2:
S 1.00
NP 0.10
VP 0.30
Astronomers
VP 0.70
PP 1.00
V 1.0
NP 0.18
P 1.0
NP 0.18
saw
stars
with
ears

概率上下文无关文法

给定一个语法分析树，我们可以计算它的概率：

$$P(T) = \prod_{i=1}^{n} P(\mathrm{RHS}_i | \mathrm{LHS}_i)$$

根据句法分析树的概率进行选择

概率上下文无关文法

如何计算每条文法规则的概率?

根据语料库训练出PCFG所需要的参数!

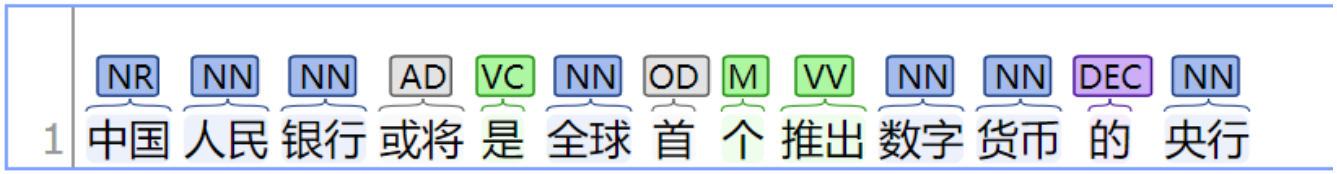
短语句法分析展示

例句：中国人民银行或将是全球首个推出数字货币的央行

https://corenlp.run

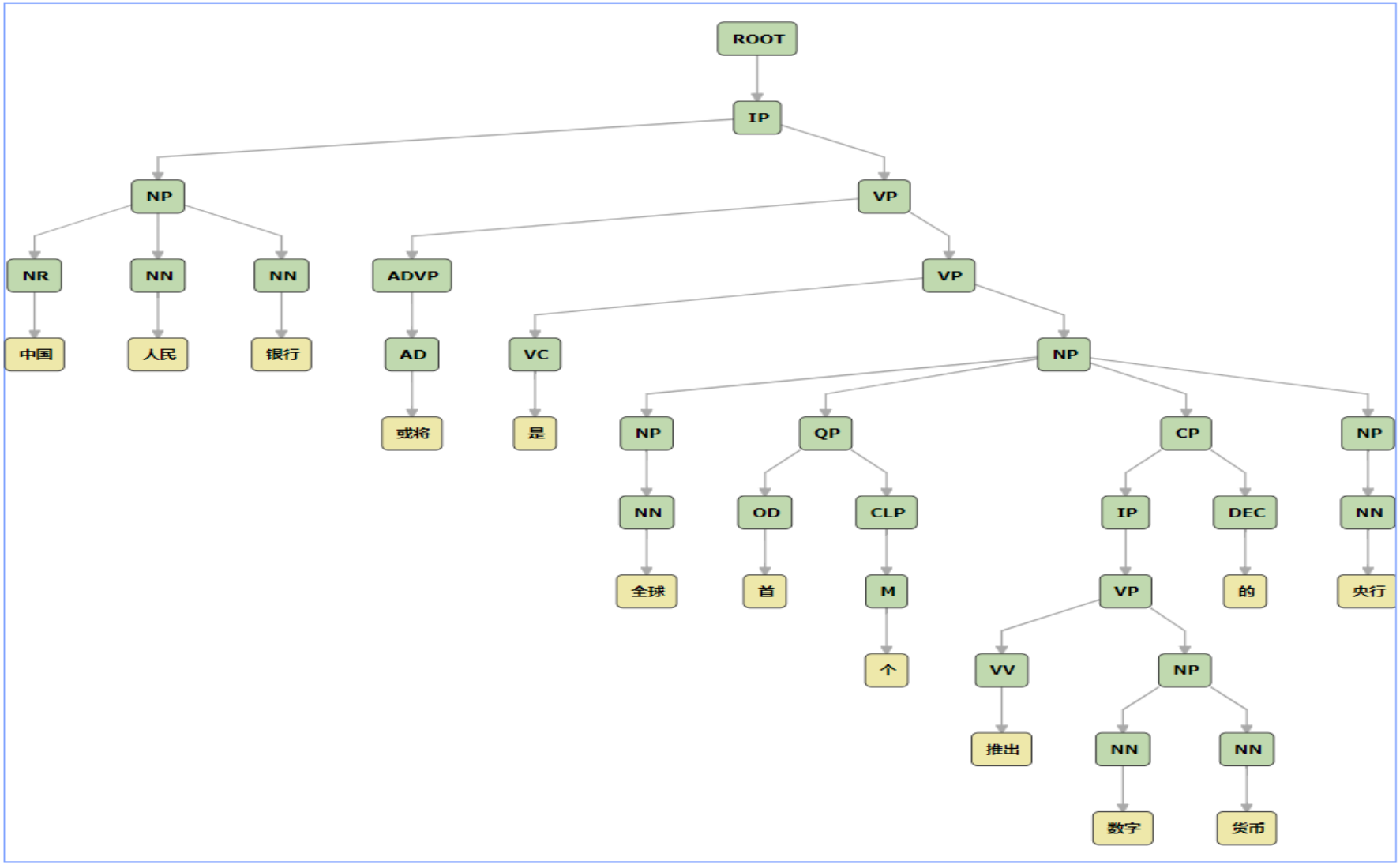
例句：中国人民银行或将是全球首个推出数字货币的央行

Part-of-Speech:



例句：中国人民银行或将是全球首个推出数字货币的央行

Constituency Parse:

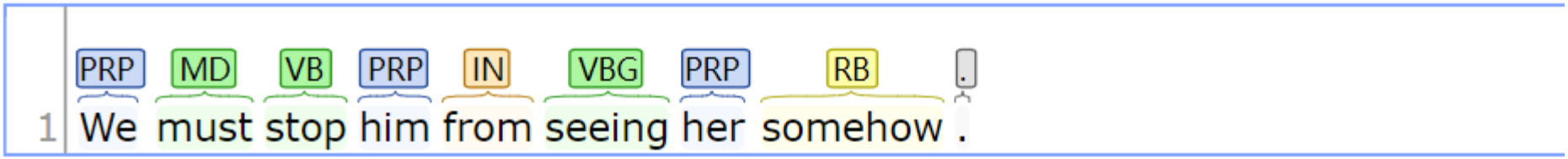


例句：We must stop him from seeing her somehow.

https://corenlp.run

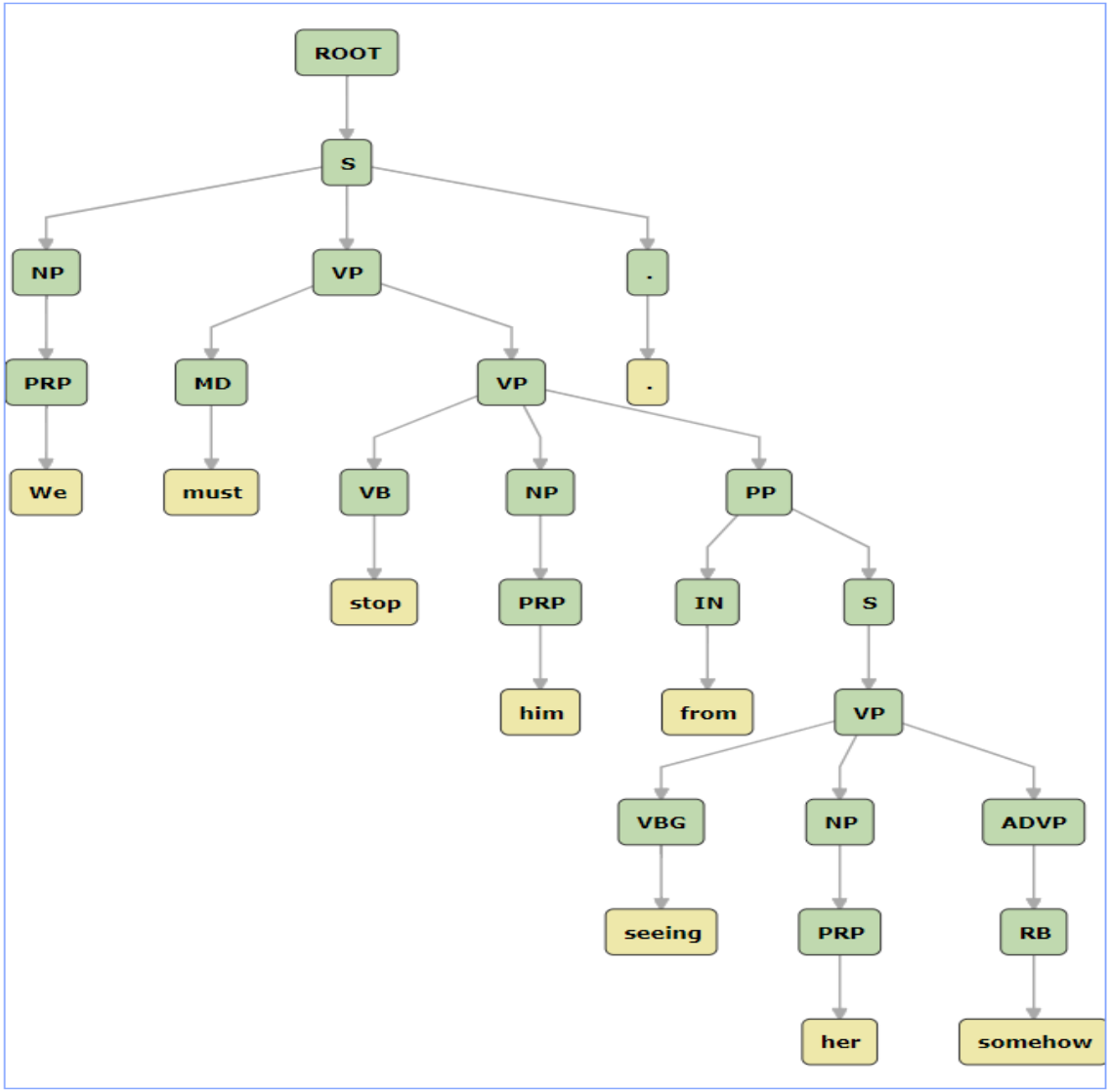
例句：We must stop him from seeing her somehow.

Part-of-Speech:



例句：We must stop him from seeing her somehow.

Constituency Parse:



句法分析

1) 短语句法分析

2) 依存句法分析

1、依存句法分析

依存句法分析

❑依存句法理论

现代依存语法理论的创立者是法国语言学家吕西安·泰尼埃 (Lucien Tesnière , 1893-1954);

泰尼埃认为：一切结构句法现象可以概括为关联(connexion)、组合(jonction)和转位(tanslation)这三大核心。句法关联建立起词与词之间的从属关系，这种从属关系是由支配词和从属词联结而成；动词是句子的中心，并支配其他成分，它本身不受其他任何成分的支配。

依存句法分析

❑依存句法理论

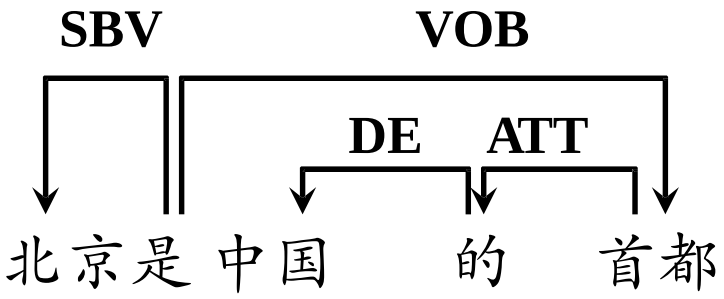
现代依存语法理论的创立者是法国语言学家吕西安·泰尼埃 (Lucien Tesnière , 1893-1954);

泰尼埃认为：一切结构句法现象可以概括为关联(connection)、组合(junction)和转位(transfer)这三大核心。句法关联建立起词与词之间的从属关系，这种从属关系是由支配词和从属词联结而成；动词是句子的中心，并支配其他成分，它本身不受其他任何成分的支配。

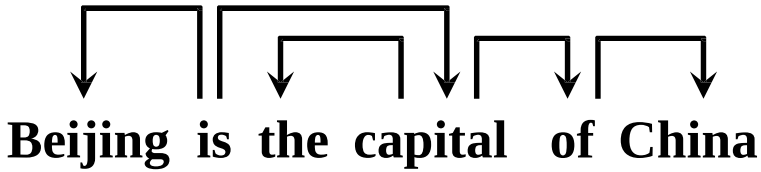
依存句法分析

在依存语法理论中，依存就是指词与词之间支配与被支配的关系，这种关系不是对等的，而是有方向的。处于支配地位的成分称为支配者(governor)，而处于被支配地位的成分称为从属者 (modifier)。

依存句法分析



(a) 有向图-1

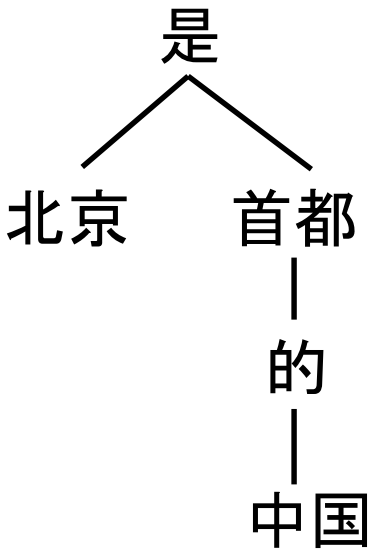


(b) 有向图-2

两个有向图用带有方向的弧(或称边)来表示两个成分之间的依存关系，支配者在有向弧的发出端，被支配者在箭头端，我们通常说被支配者依存于支配者。

依存句法分析

(c) 依存树



图(c)是用树表示的依存结构，树中子节点依存于该节点的父节点。

依存句法分析

1970年计算语言学家J. Robinson在论文《依存结构和转换规则》中提出了依存语法的4条公理：

(1) 一个句子只有一个独立的成分；

(2) 句子的其他成分都从属于某一成分；

(3) 任何一成分都不能依存于两个或多个成分；

(4) 如果成分A直接从属于成分B，而成分C在句子中位于A和B之间，那么，成分C或者从属于A，或者从属于B，或者从属于A和B之间的某一成分。

依存句法分析

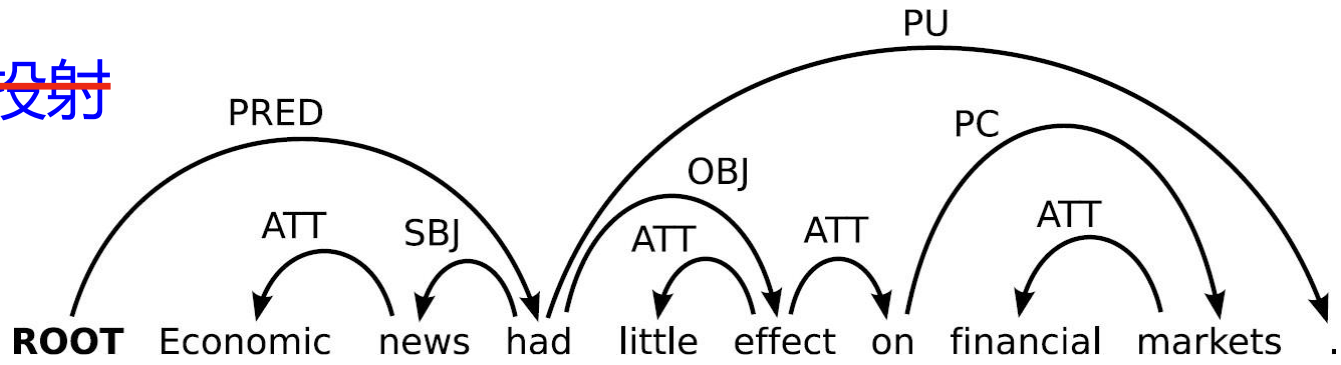
这4条公理相当于对依存图和依存树的形式约束为：

- 单一父结点(single headed)
- 连通(connective)
- 无环(acyclic)
- 可投射(projective)

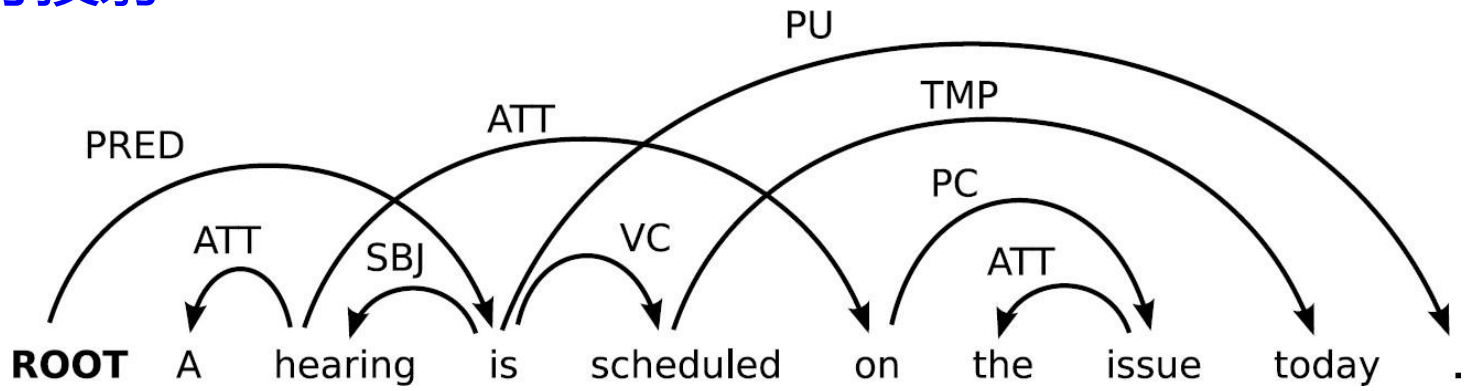
由此来保证句子的依存分析结果是一棵有根的树结构

依存句法分析

~~可投射~~



~~非可投射~~



依存句法分析

❑ 依存语法的优势

1) 依存关系和实际的语义关系比较接近，有助于对句子的语义方面的理解；

2) 定义相对比较简单，有助于高效率的句法分析；

3) 因为能够有效建模长距离依赖关系，依存句法更适合词序列比较自由、灵活的语言；

依存句法分析

❑ 依存句法分析方法

依存句法分(dependency parsing)的任务就是分析出句子中所有词汇之间的依存关系。

依存句法分析

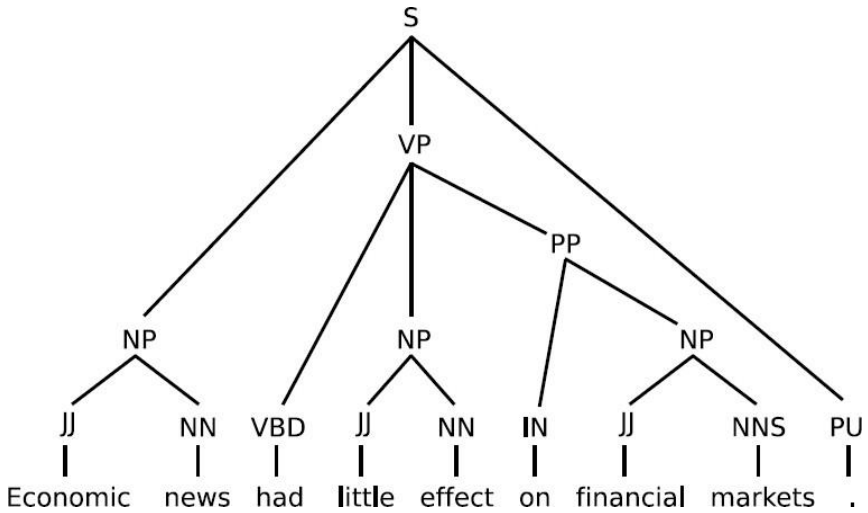
句法分析算法可大致归为以下4类：

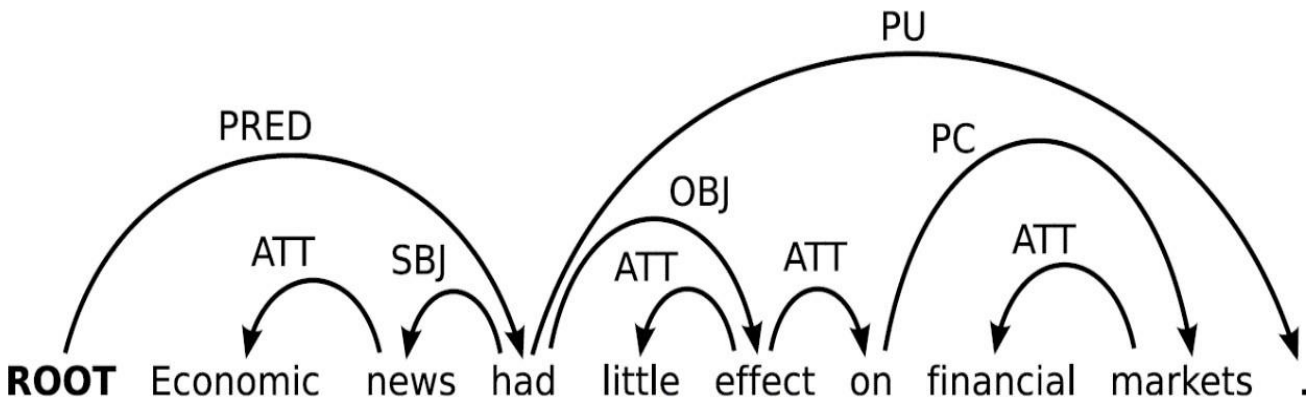
- 生成式的分析方法(generative parsing)
- 判别式的分析方法(discriminative parsing)
- 决策式的(确定性的)分析方法(deterministic parsing)
- 基于约束满足的分析方法(constraint satisfaction parsing)

2、短语结构与依存结构的关系

短语结构与依存结构

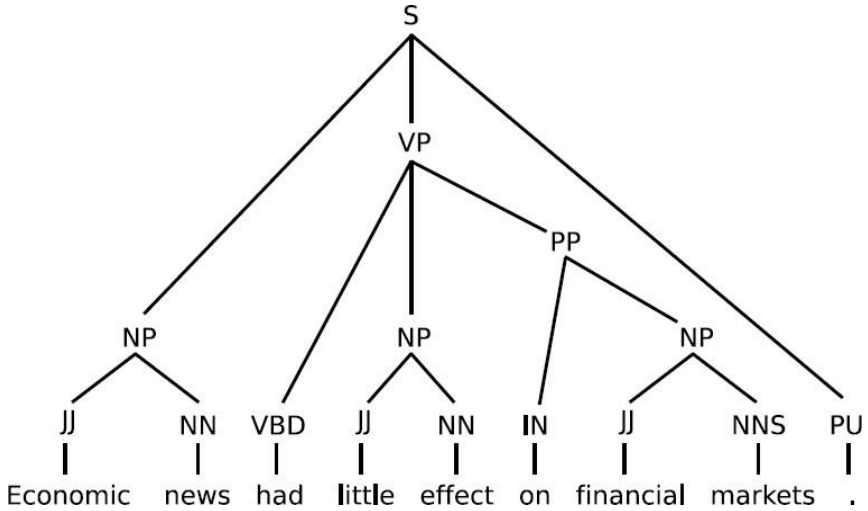
依存结构表达的信息和短语结构句法树不一样，可以表达**更长距离**的信息依存关系

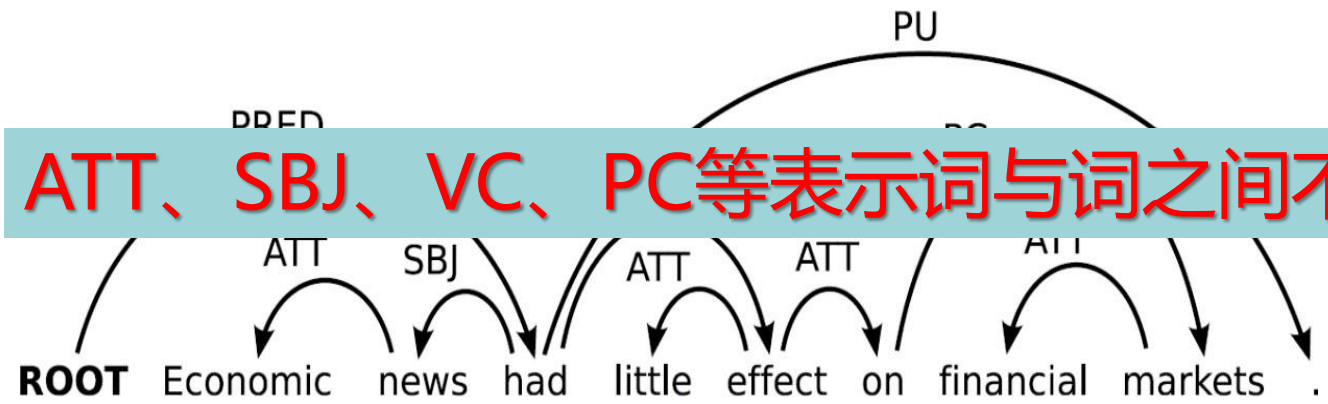




短语结构与依存结构

依存结构表达的信息和短语结构句法树不一样，可以表达**更长距离**的信息依存关系





ATT、SBJ、VC、PC等表示词与词之间不同的依存关系

短语结构与依存结构的关系

- 短语结构可转换为依存结构
- 实现方法：

(1) 定义中心词抽取规则，产生中心词表；

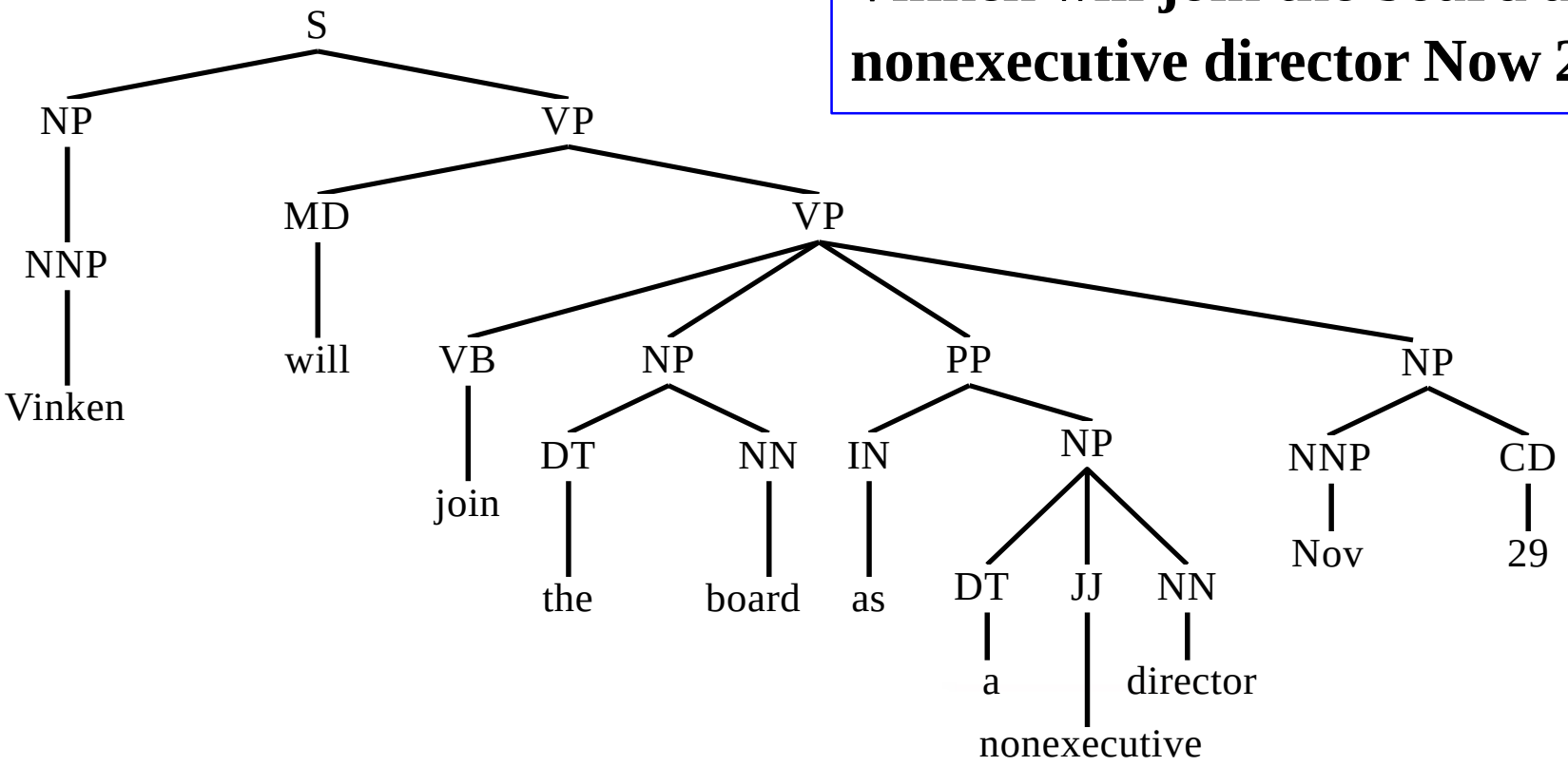
(2) 根据中心词表，为每个节点选择中心子节点；

(3) 将非中心子节点的中心词依存到中心子节点的中心词上，得到相应的依存结构。

短语结构与依存结构的关系

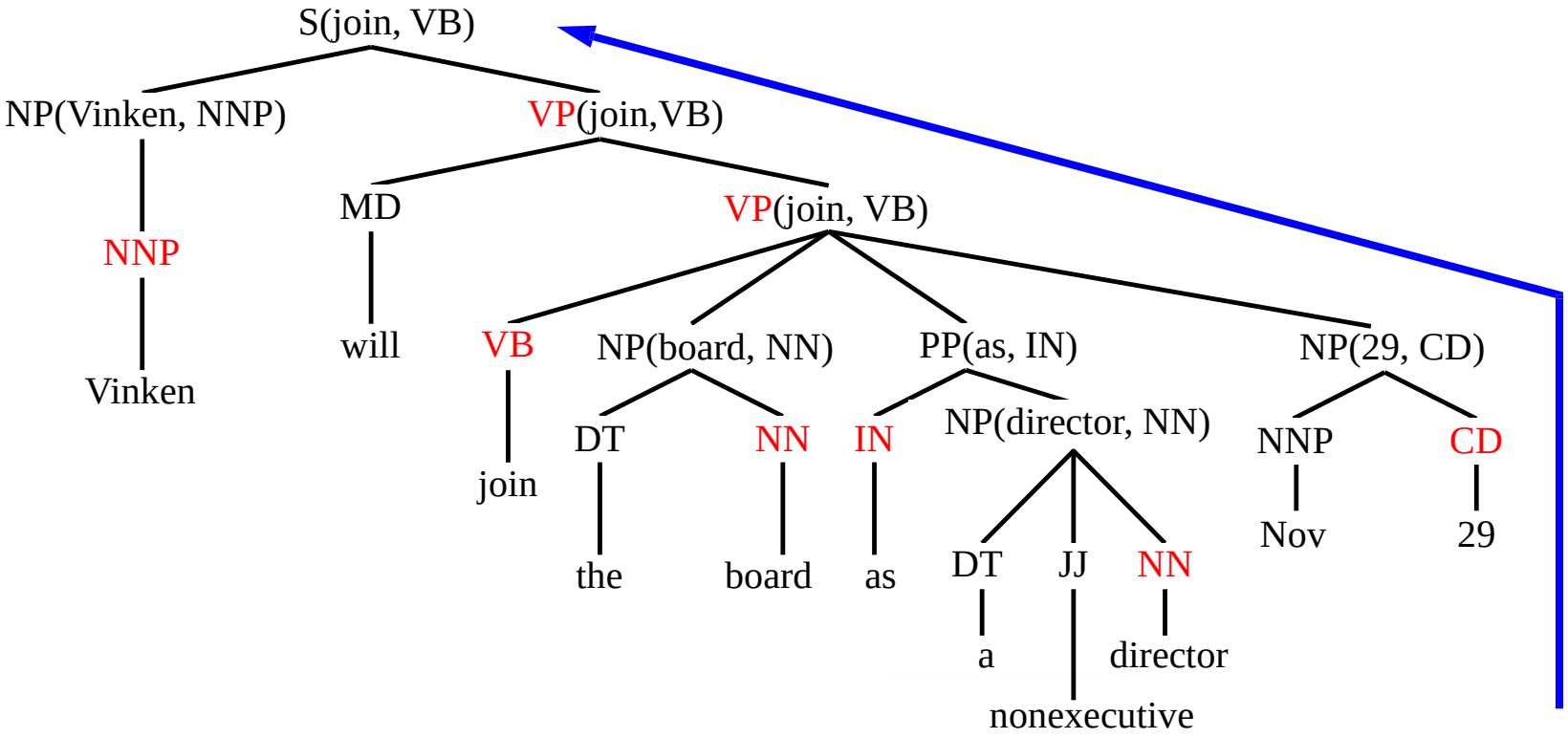
例如：给定如下短语结构树

Vinken will join the board as a nonexecutive director Now 29



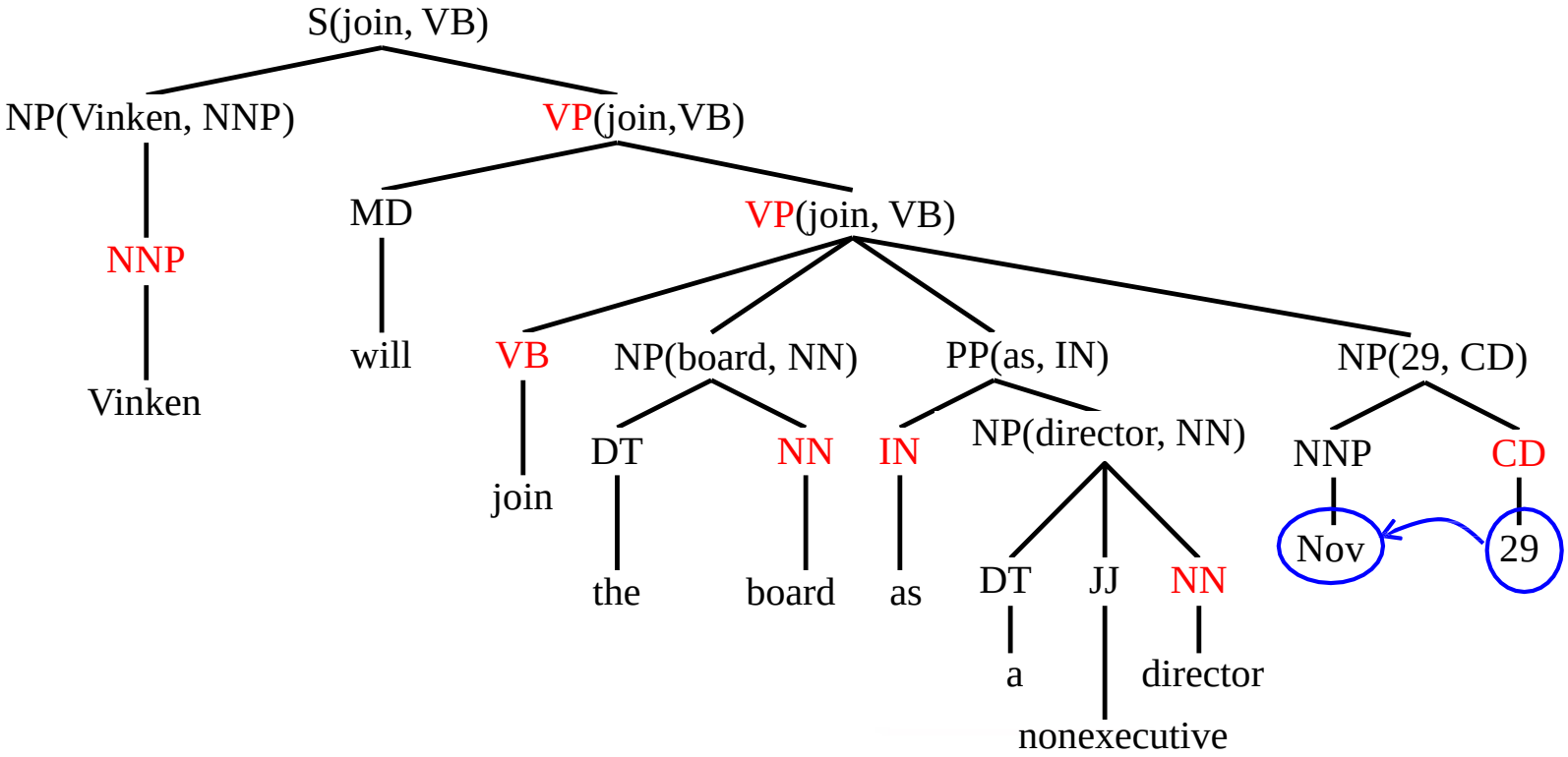
短语结构与依存结构的关系

- 根据中心词表为每个节点选择中心子节点(中心词通过自底向上传递得到)



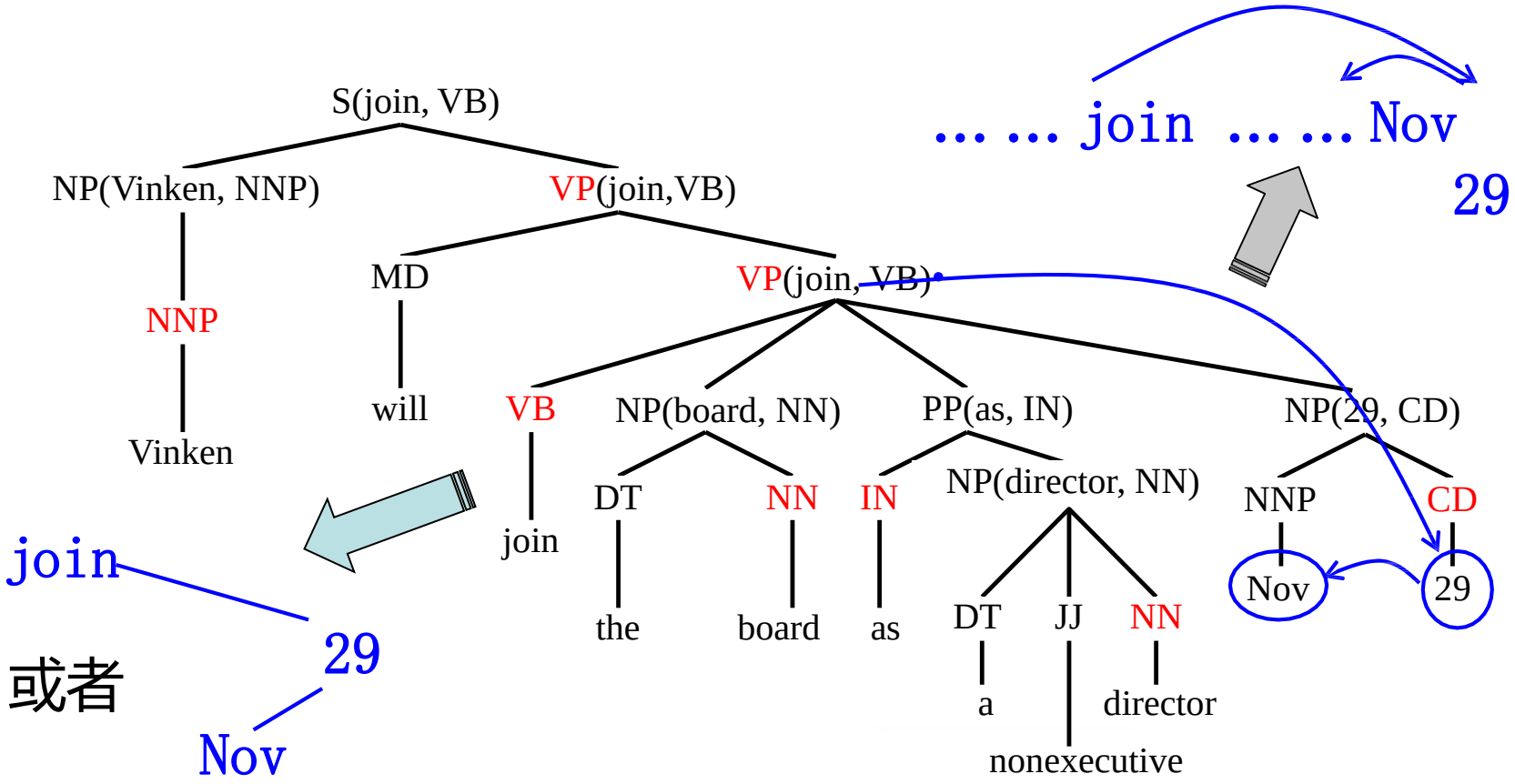
短语结构与依存结构的关系

- 将非中心子节点的中心词依存到中心子节点的中心词上



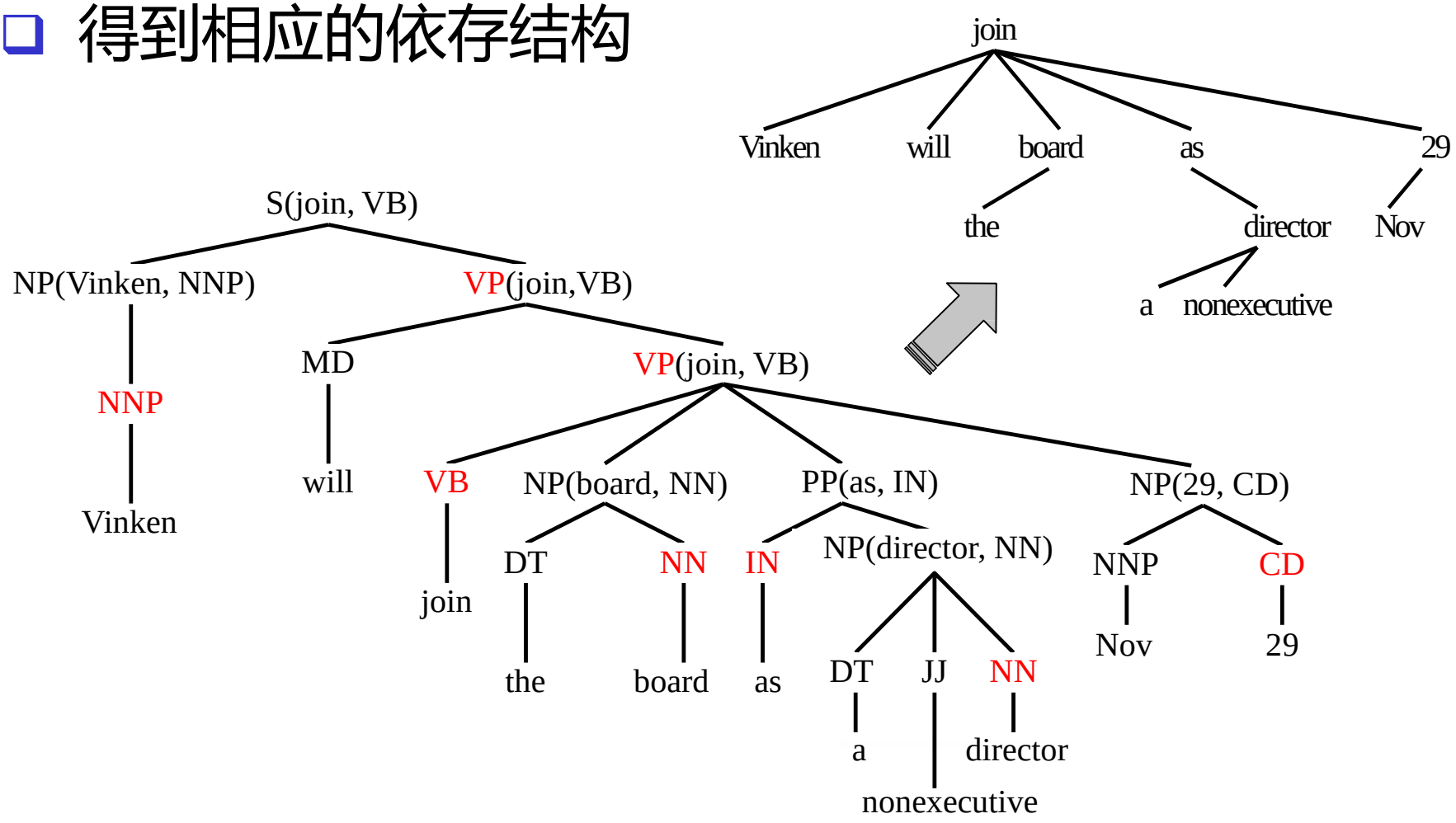
短语结构与依存结构的关系

- 将非中心子节点的中心词依存到中心子节点的中心词上



短语结构与依存结构的关系

- 得到相应的依存结构



思考

短语结构→依存结构 ✔

依存结构→短语结构 ???

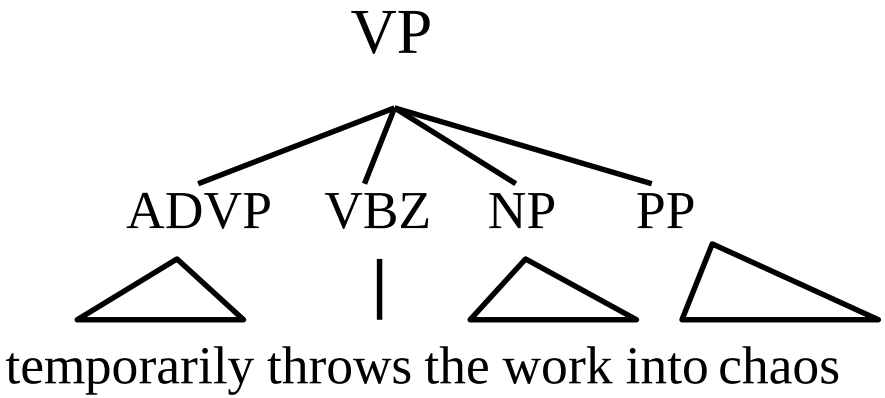
3、汉英句法结构特点对比

汉英句法结构特点对比

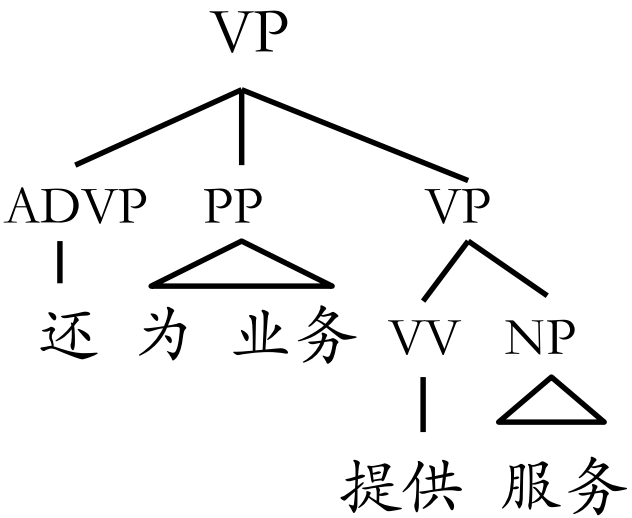
说明：撇开汉语的分词问题和词性消歧错误可能对句法分析器带来的影响，即保证句法分析器的输入为完全正确的词性序列，仅仅考虑句子结构本身的问题；

汉英句法结构特点对比

(1)英语短语绝大多数以左部为中心，而汉语短语比较复杂，大多数短语类是以右部为短语中心，除了动词和介词的补语在它们的中心词之后。如：



(a)



(b)

汉英句法结构特点对比

在例图(a)中，介词短语 into chaos 在动词 throw 的右边，而在图(b)所示的汉语例子中恰好相反，介词短语“为业务”在动词前面。这种差异意味着在英语句子中附加在动词后面的补语引起的歧义是句法分析器需要解决的主要问题，而在汉语句子中很少有这种歧义存在。

汉英句法结构特点对比

(2) 在汉语句子中没有做主语的先行代词的情况普遍存在，但在英语中这种情况很少出现。这样就使得汉语句法分析器很难判断一个输入到底是没有主语的子句结构还是仅仅是一个动词短语VP， 如：

He thinks it is true. / 他认为□是对的。

汉英句法结构特点对比

英语中当多个单句连接起来构成复句的时候，单句与单句之间需要有显式的连接词或者短语。汉语则不同，一个句子是表达一个完整意义的语言单元，这种特点在长句中表现得特别明显。

汉英句法结构特点对比

这些长句内部的各个简单句是为了表意的需要而连接在一起的，它们彼此的句法结构完全是独立的，表示彼此之间逻辑关系的连接词不是必需的，这类长句在汉语中称之为“流水复句”，例如：

“我现已步入中年，每天挤车，搞得我精疲力尽，这种状况，直接影响我的工作，家里的孩子也没人照顾。”

汉英句法结构特点对比

- 汉语长句的层次化句法分析方法

(1) 对包含“分割”标点的长句进行分割;

(2) 对分割后的各个子句分别进行句法分析(第一 级分析)，分析得到的子树根节点的词类或者短语类别标记作为第二级句法分析的输入;

(3) 通过第二遍分析找到各子句或短语之间的结构关系，从而获得最终整句的最大概率分析树。

思考题

1. 什么是CFG/PCFG?
2. 简述依存句法与CFG/PCFG的区别?
3. 何为依存句法树的投射性?

Thank you!

权小军 中山大学数据科学与计算机学院