

1

第5章 句法分析(III)

目录

2

1. 依存句法分析
2. 汉英句法结构特点对比
3. 局部句法分析
4. 小结

3

1 依存句法分析

1 依存句法分析

4

□ 依存句法理论

- ▣ Dependency Grammar
- ▣ 法国语言学家吕西安·泰尼埃(Lucien Tesnière, 1893-1954)
- ▣ 1953年出版: 《结构句法概要》

1 依存句法分析

5

L. Tesnière 理论认为

- 一切结构句法现象可以概括为**关联**(connexion)、**组合**(jonction)和**转位**(translation)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。
- 欧洲传统的语言学突出一个句子中**主语**的地位，句中其它成分称为“**谓语**”。依存语法打破了这种主谓关系，认为“谓语”中的动词是一个句子的中心，其他成分与动词直接或间接地产生联系。

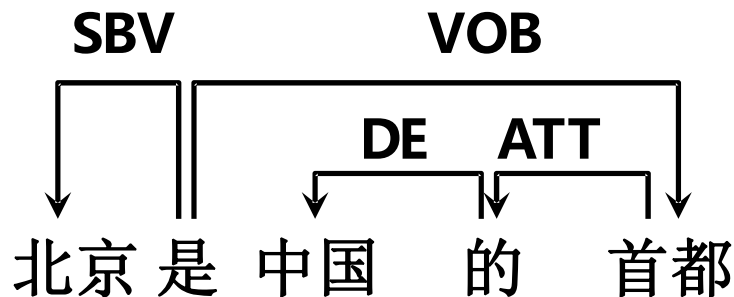
1 依存句法分析

6

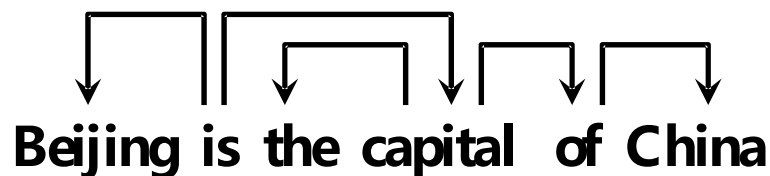
- 依存语法理论中，“依存”就是指词与词之间支配与被支配的关系，这种关系不是对等的，而是有方向的。
- 处于支配地位的成分称为支配者(head)，而处于被支配地位的成分称为从属者(dependent)。

1 依存句法分析

7



(e) 有向图-1

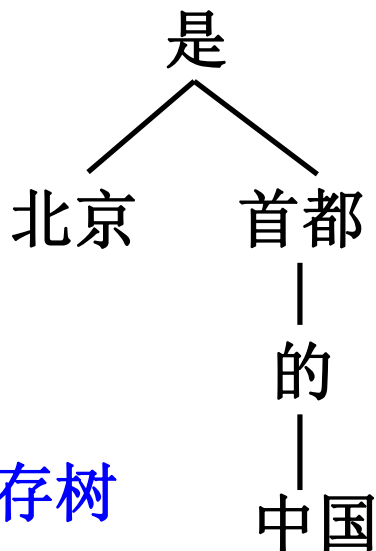


(e) 有向图-2

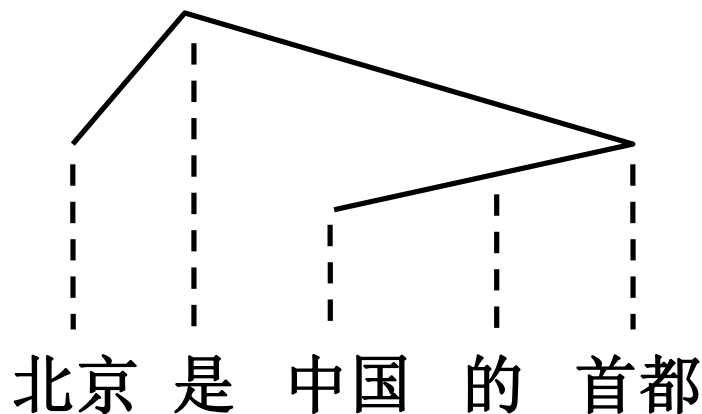
- 两个有向图用带有方向的弧(或称边, arc)来表示两个成分之间的依存关系, 支配者在有向弧的发出端, 从属者在箭头端, 我们通常说从属者依存于支配者。

1 依存句法分析

8



(f) 依存树



(g) 依存投射树

- 图(f)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。
- 图(g)是带有投射线的树结构, 实线表示依存联结关系, 位置低的成份依存于位置高的成份, 虚线为投射线。

1 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四条公理：
 - (1) 一个句子只有一个独立的成分（不从属于任何成分）；
 - (2) 句子的所有其他成分都从属于某一成分；
 - (3) 任何一个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。

1 依存句法分析

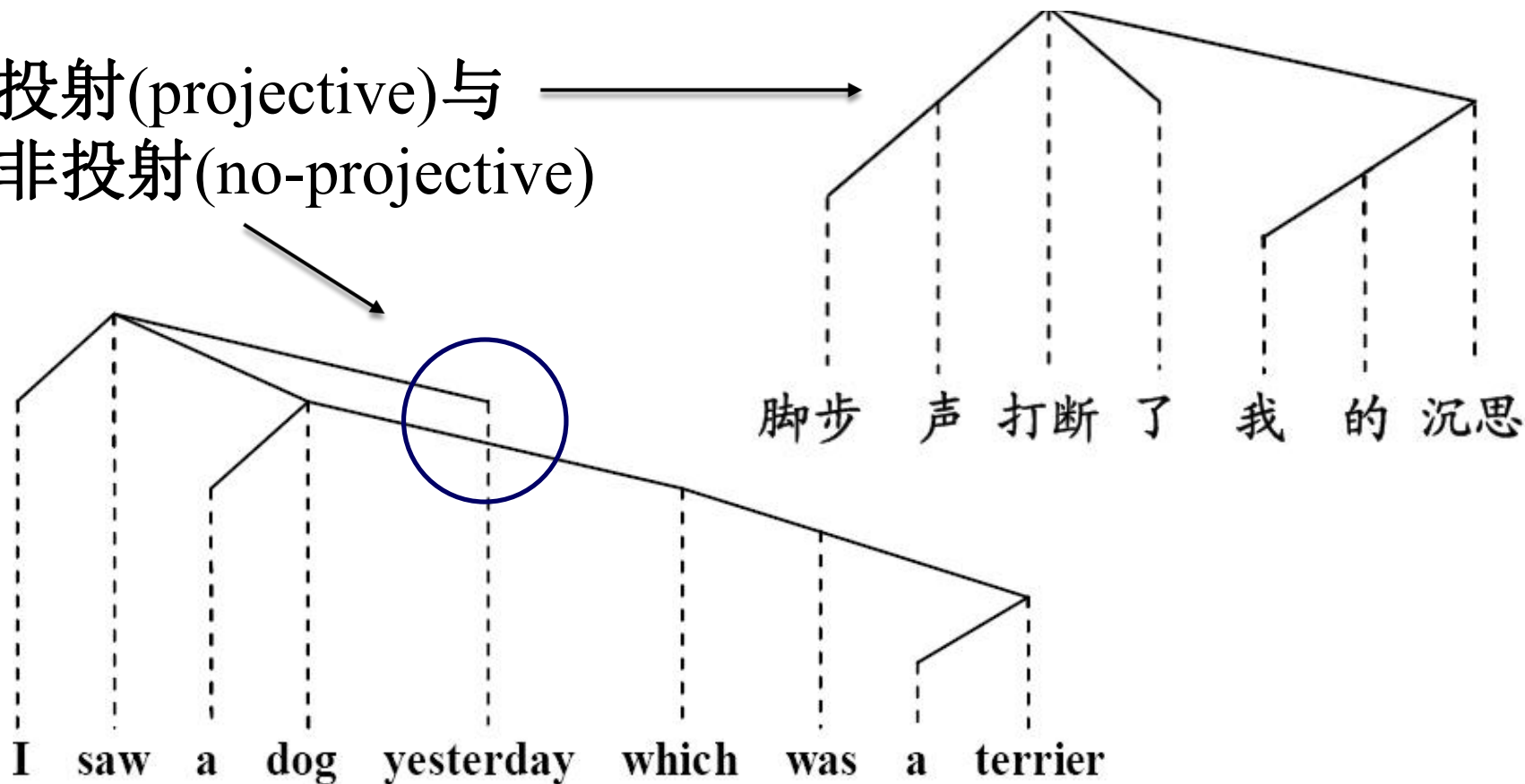
10

- 这四条公理相当于对依存图或依存树的形式约束为：
 - ▣ 单一父结点(single head, 加一个特殊的root结点来保证)
 - ▣ 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(root)”的树结构。

1 依存句法分析

11

投射(projective)与
非投射(no-projective)



1 依存句法分析

12

- 但是依存结构理论是允许非投射结构存在的
- 有时，我们不能通过投射结构来完整的表达句子成分之间的依存关系



1 依存句法分析

13

□ 依存句法分析方法

- ▣ 依存句法分析(dependency parsing)的任务就是分析出句子中所有词汇之间的依存关系。
- ▣ 建立一个依存句法分析器一般需要完成以下三部分工作：
 - (1) 依存句法结构描述
 - (2) 分析算法设计与实现
 - (3) 文法规则或参数学习

1 依存句法分析

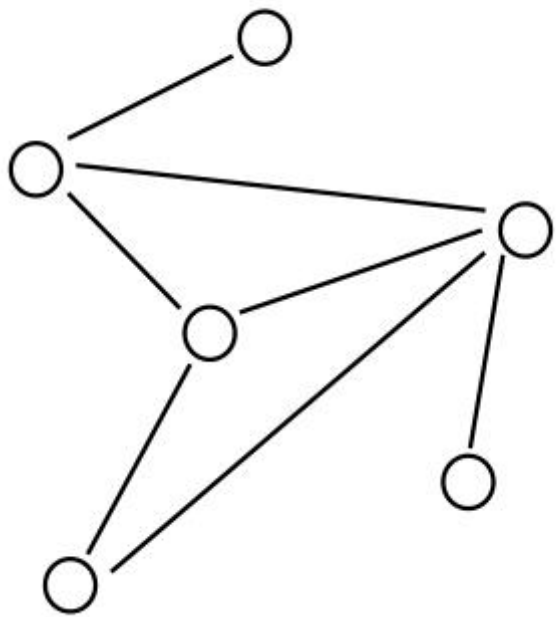
14

- 目前依存句法结构描述一般采用有向图方法或依存树方法，所采用的句法分析算法可大致归为以下四类：
 - ▣ 基于语法的
 - 上下文无关依存语法 (context-free dependency grammar)
 - 有限制依存语法 (constraint dependency grammar, <https://cl.lingfil.uu.se/~nivre/docs/05133.pdf>)
 - ▣ 数据驱动的
 - 基于图的 (graph-based)
 - 基于转换的 (transition-based)

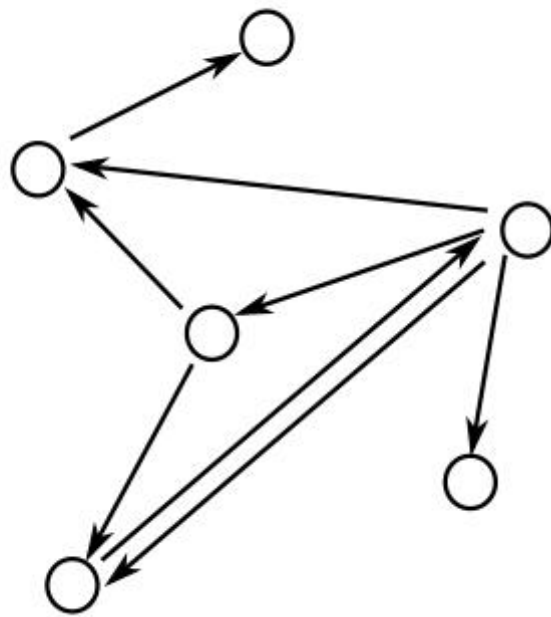
数据驱动-基于图的方法

15

无向图 (undirected graph)



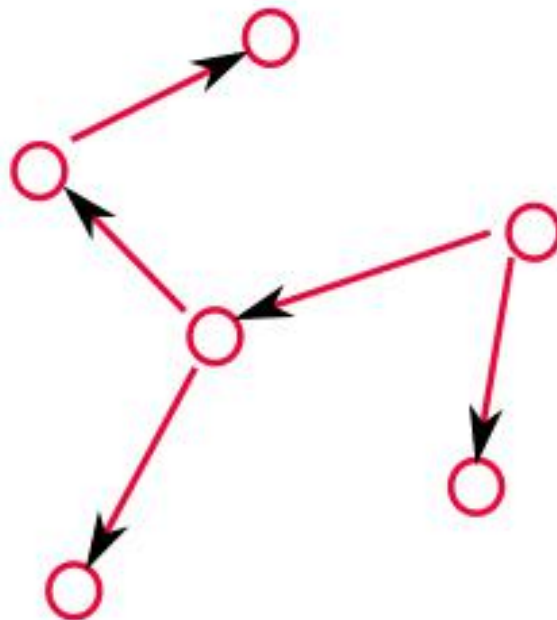
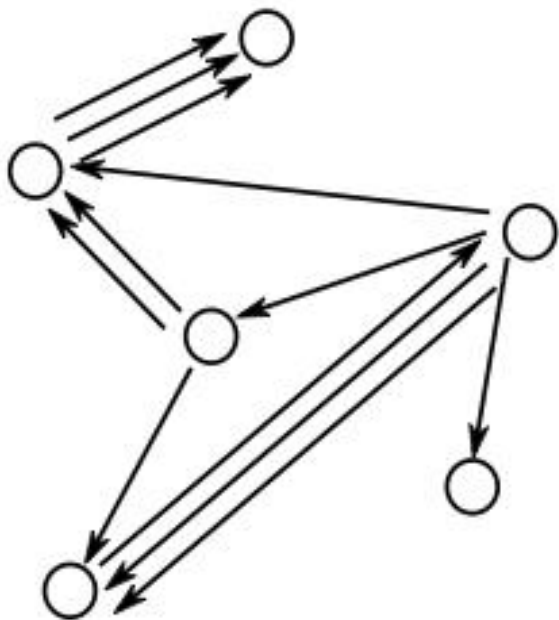
有向图 (directed graph, digraph)



数据驱动-基于图的方法

16

多边有向图 (multi-digraph) \longrightarrow 有向生成树 (directed spanning tree)



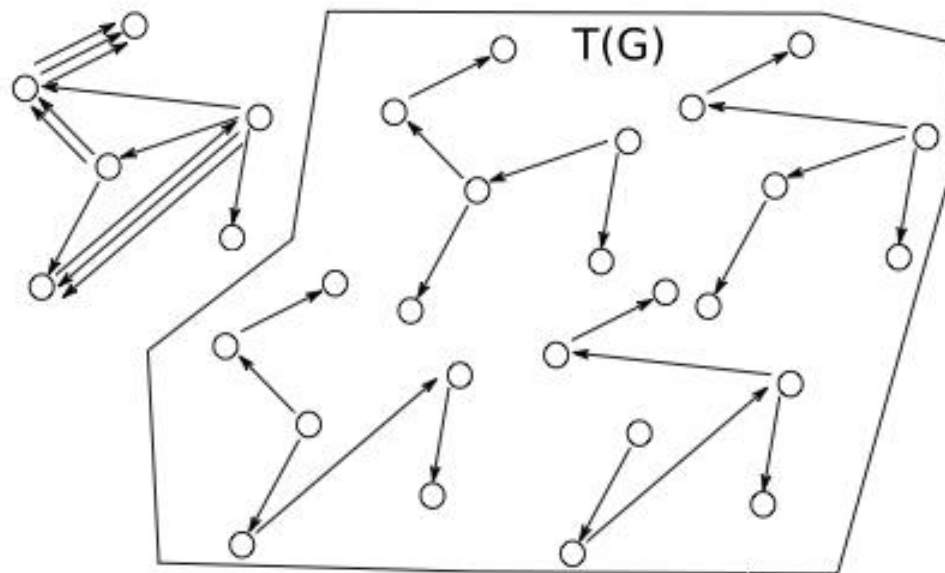
用最少的边把原图的节点连起来形成一棵树：
节点数不变，边数为结节点数少1

数据驱动-基于图的方法

17

假设我们的多边有向图的每条边是有权重的，那么我们可以对应定义一个有权有向生成树（**weighted directed spanning tree**）。并定义一个有权有向生成树的总权重是其所有边得权重之和。

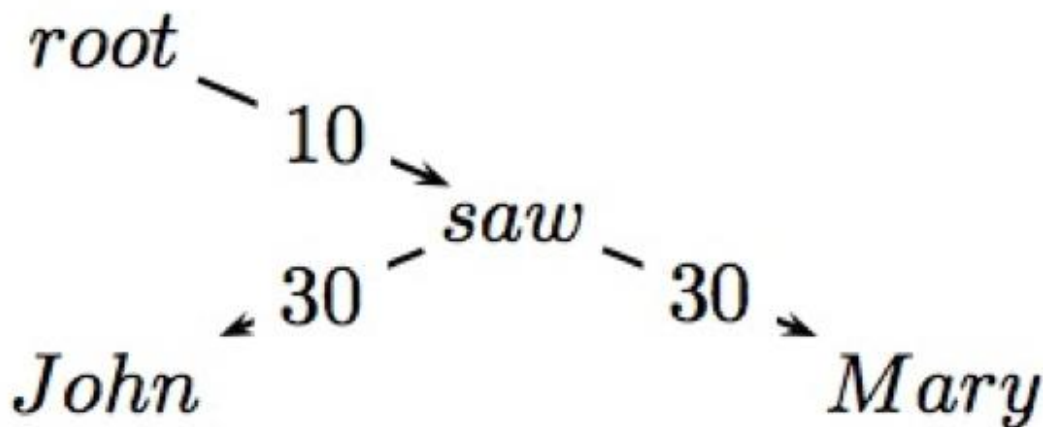
定义 $T(G)$ 为一个multi-digraph所有可能的**weighted directed spanning tree**的集合，那么我们可以定义一个最大（权重）生成树（**max spanning tree**）的问题。



数据驱动-基于图的方法

18

自然地，我们可以想到依存解析也可以变为一个最大生成树的问题（因为每一个生成树都是一个合法的依存解析树），其中，节点为词语，词语之间的多边为所有可能存在的依存关系，每条边的权重为模型预测一个节点与另外一个节点之间存在某种依存关系的数值。



数据驱动-基于图的方法

19

Features? 节点与节点之间的权重该用什么特征来计算？

McDonald et al. [2005] 介绍了一系列的特征。例如，词语的词性和可能存在的联系。开始使用神经网络后，对于特征的依赖逐渐减小 <https://arxiv.org/abs/1611.01734>

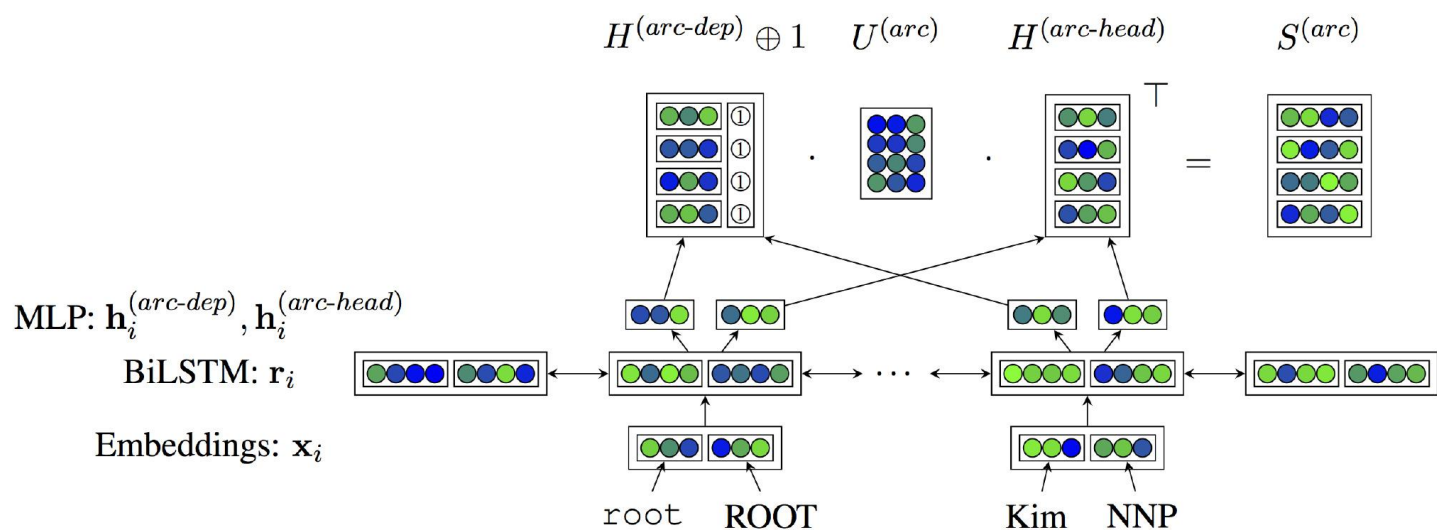


Figure 1: BiLSTM with deep biaffine attention to score each possible head for each dependent, applied to the sentence “Casey hugged Kim”. We reverse the order of the biaffine transformation here for clarity.

数据驱动-基于图的方法

20

接下来，学习问题就变得简单了：

- 利用模型推断最大生成树
- 与真实的依存树计算损失更新权重

怎么推断？

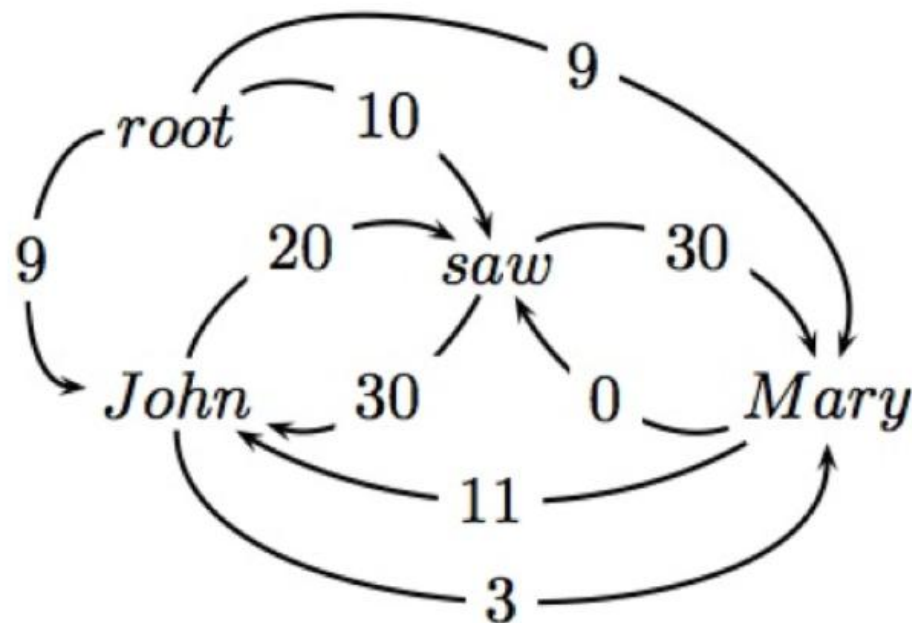
Dynamic programming, [Eisner 1996], projective, $O(n^3)$

Chu-Liu-Edmonds Algorithm, Developed Chu and Liu [1965] and Edmonds [1967], non-projective, $O(n^2)$

数据驱动-基于图的方法

21

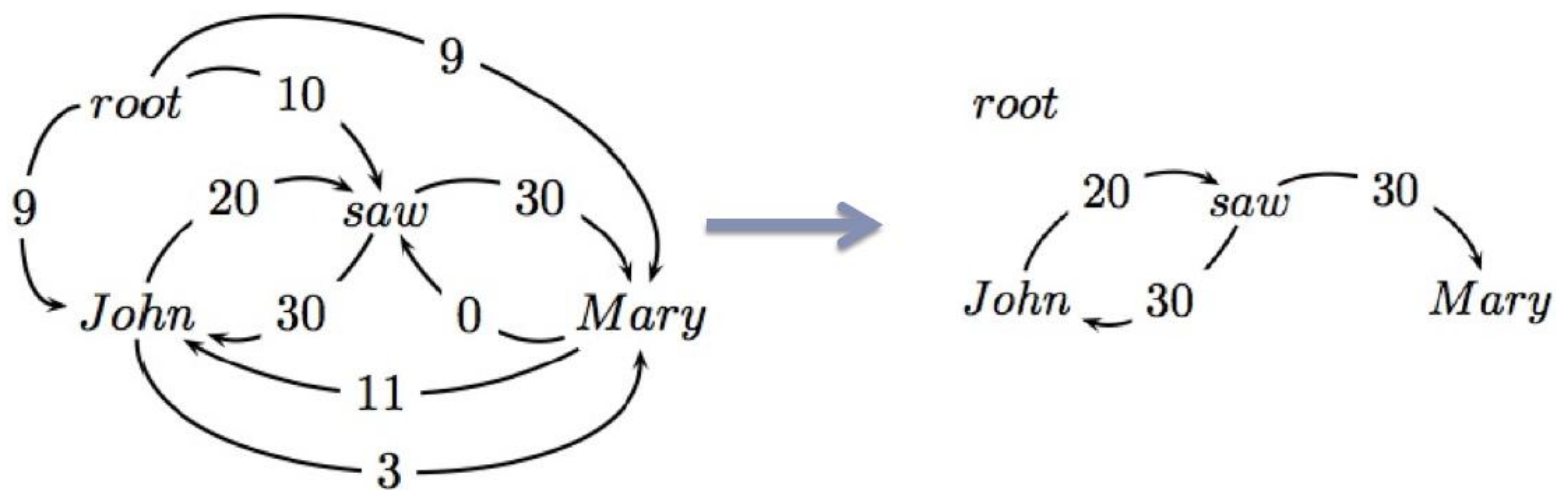
- ▶ $x = \text{root John saw Mary}$
- ▶ Remove all arcs into the root node



数据驱动-基于图的方法

22

- Find highest scoring incoming arc for each vertex

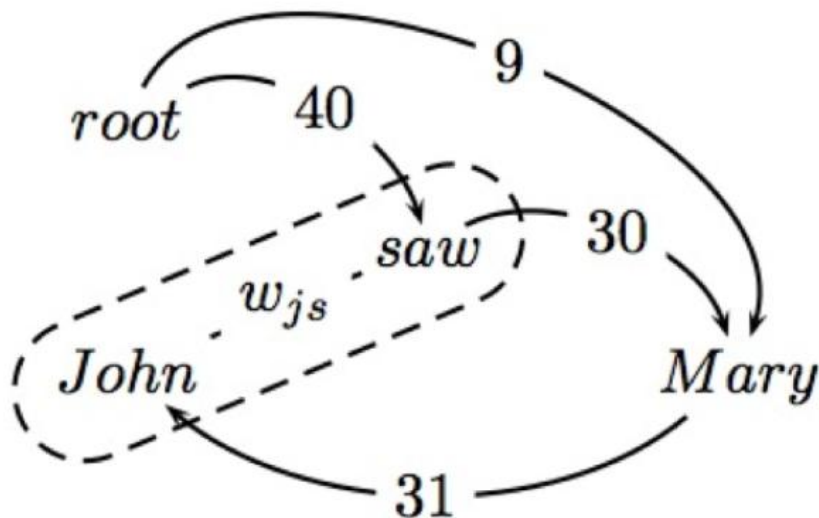


- If this is a tree, then we have found an MST

数据驱动-基于图的方法

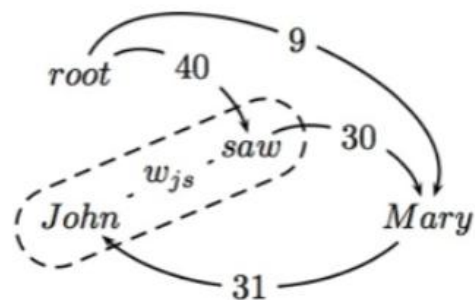
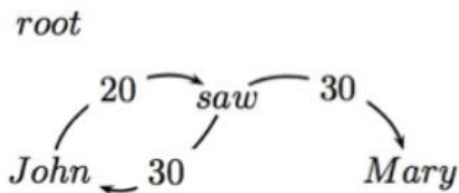
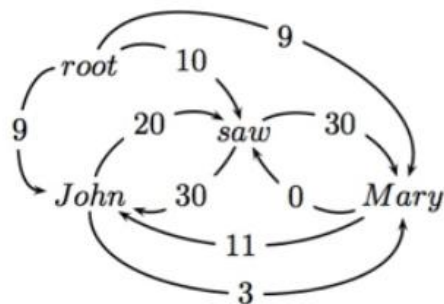
23

- ▶ If not a tree, identify cycle and contract
- ▶ Recalculate arcs weights into and out-of the cycle



数据驱动-基于图的方法

24

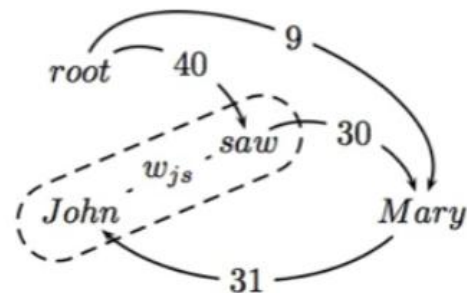
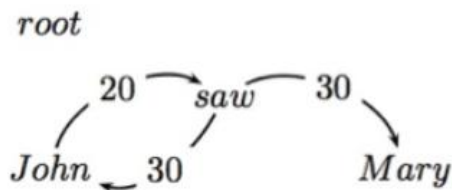
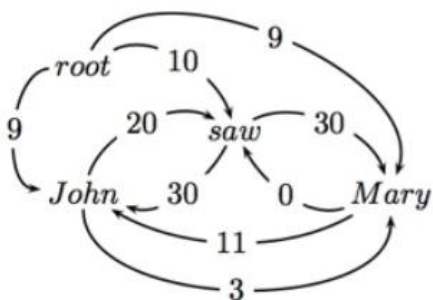


► Outgoing arc weights

- Equal to the max of outgoing arcs over all vertices in the cycle
- e.g., John \rightarrow Mary is 3, and saw \rightarrow Mary is 30

数据驱动-基于图的方法

25



► Incoming arc weights

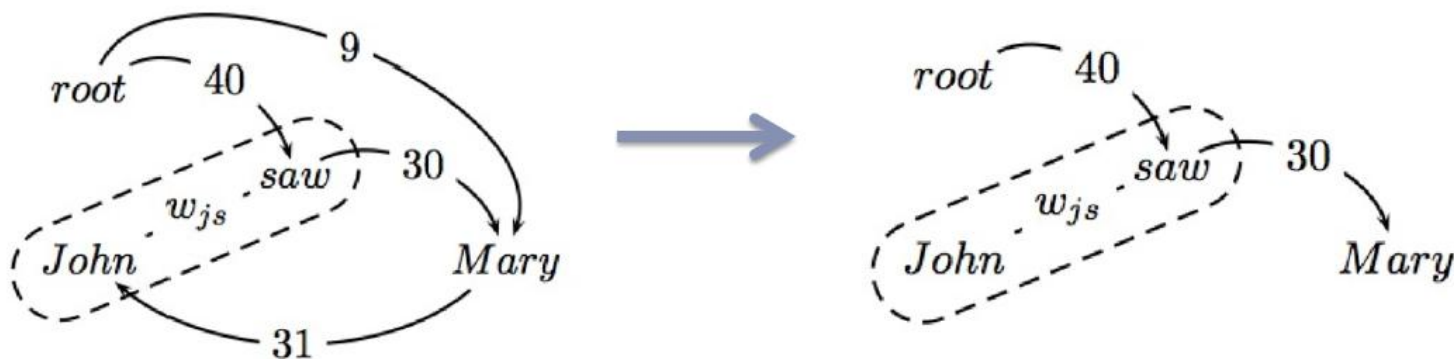
- Equal to the weight of best spanning tree that includes head of incoming arc, and all nodes in cycle
- $\text{root} \rightarrow \text{saw} \rightarrow \text{John}$ is 40
- $\text{root} \rightarrow \text{John} \rightarrow \text{saw}$ is 29

It is easy to prove that we need to consider **only edges from the cycle** and an arc from the considered word to the cycle

数据驱动-基于图的方法

26

This is a tree and the MST for the contracted graph!

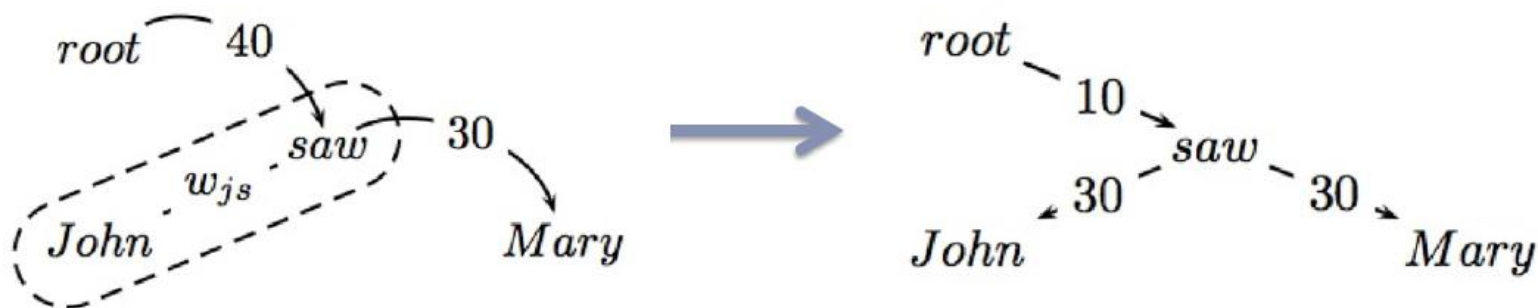


- ▶ Go back (in recursion) and reconstruct the final graph

数据驱动-基于图的方法

27

- ▶ This the MST



数据驱动-基于图的方法

- (1) 依存句法结构描述
 - 预测预测结构变为从最大生成树集合中预测最大生成树
- (2) 分析算法设计与实现
 - 模型提出一个最大生成树集合
 - 利用算法抽取出最大生成树
- (3) 文法规则或参数学习
 - 和真实的最大生成树做损失

数据驱动-基于转移的方法

29

一个依存树可以被转化为一个动作序列，然后一个模型因此只需要去预测一个序列。更多直觉性的东西留到编译原理。

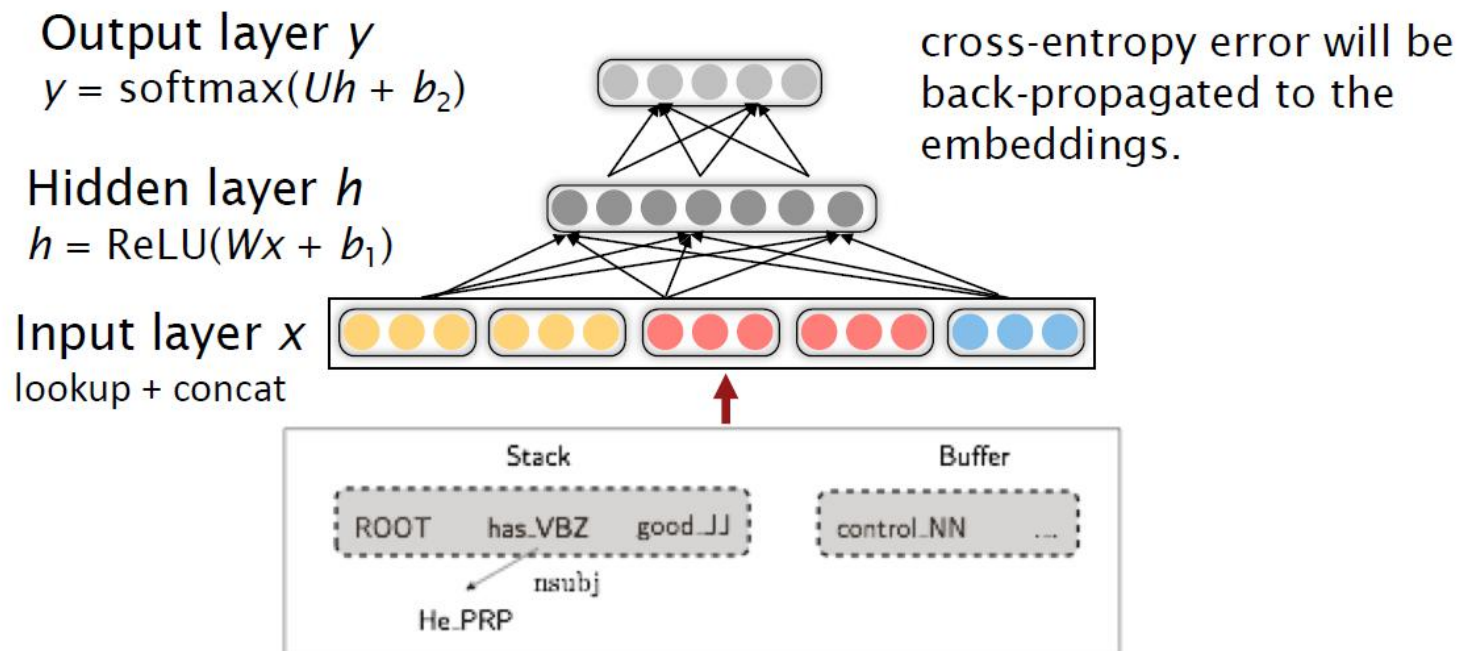
数据驱动-基于转移的方法

30

很容易地可以被拓展称为基于神经网络的版，

<https://nlp.stanford.edu/pubs/emnlp2014-depparser.pdf>

Softmax probabilities



数据驱动-基于转移的方法

31

transition-based parsing, shift-reduce algorithm [Nivre 2003], **arc-eager algorithm** [Nivre 2004], projective, $O(n)$

- The parser does a sequence of bottom up actions
 - Roughly like “shift” or “reduce” in a shift-reduce parser, but the “reduce” actions are specialized to create dependencies with head on left or right
- The parser has:
 - a stack σ , written with top to the right
 - which starts with the ROOT symbol
 - a buffer β , written with top to the left
 - which starts with the input sentence
 - a set of dependency arcs A
 - which starts off empty
 - a set of actions

数据驱动-基于转移的方法

32

动作集合:

Left-arc: 右词 (**buffer head**) 为左词 (**stack top**) 的子节点, 把**stack top**的词移出

Right-arc: 左词 (**stack top**) 为右词 (**buffer head**) 的子节点, 从**buffer**移入**stack**

Shift: 从**buffer**移入**stack**

Reduce: 把**stack top**的词移出

数据驱动-基于转移的方法

33

$[\text{root}_0]_\sigma [\text{Economic}_1 \text{ news}_2 \text{ had}_3 \text{ little}_4 \text{ effect}_5 \text{ on}_6 \text{ financial}_7 \text{ markets}_8 \text{ .9}]_\beta$

数据驱动-基于转移的方法

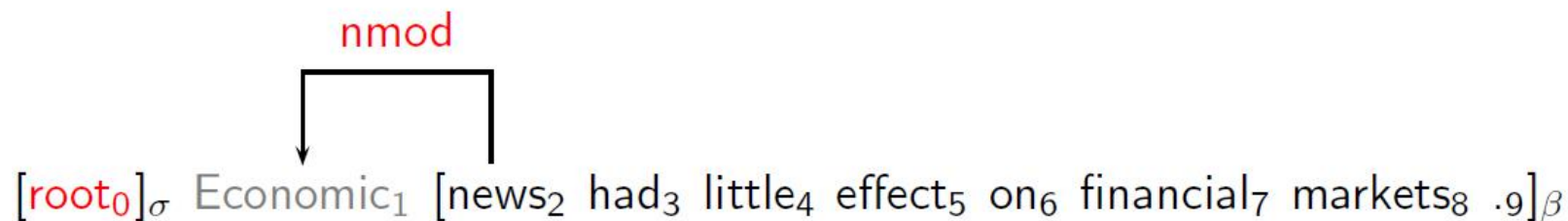
34

$[\text{root}_0 \text{ Economic}_1]_\sigma [\text{news}_2 \text{ had}_3 \text{ little}_4 \text{ effect}_5 \text{ on}_6 \text{ financial}_7 \text{ markets}_8 \text{ .9}]_\beta$

Shift

数据驱动-基于转移的方法

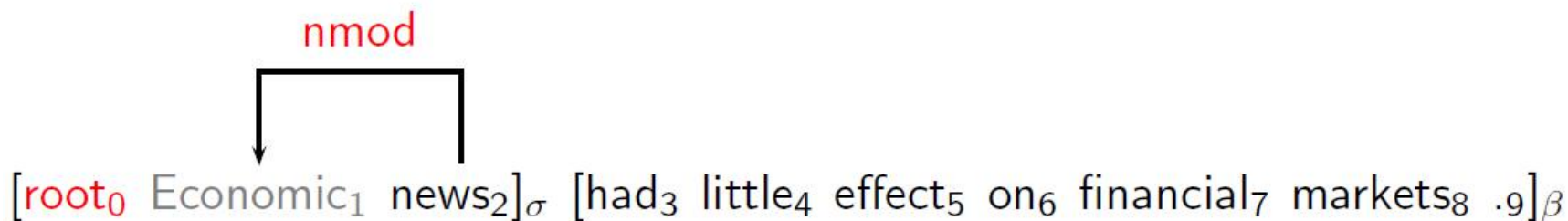
35



Left-Arc_{nmod}

数据驱动-基于转移的方法

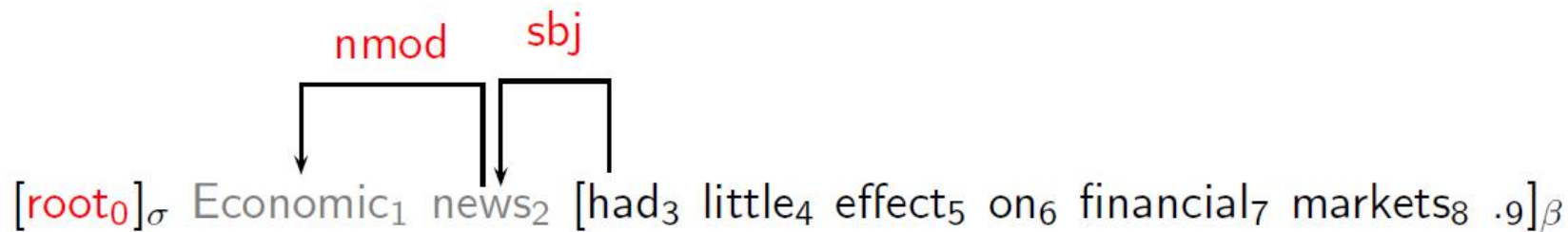
36



Shift

数据驱动-基于转移的方法

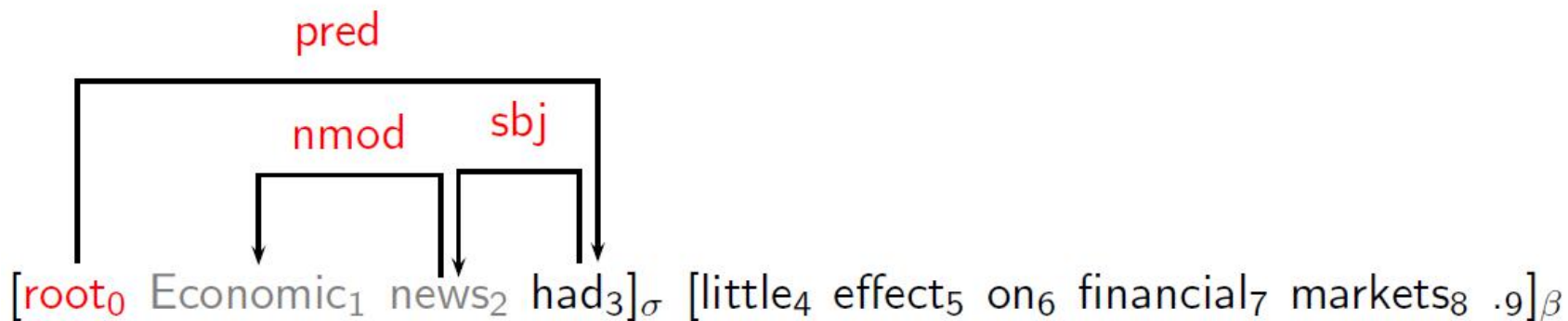
37



Left-Arc_{sbj}

数据驱动-基于转移的方法

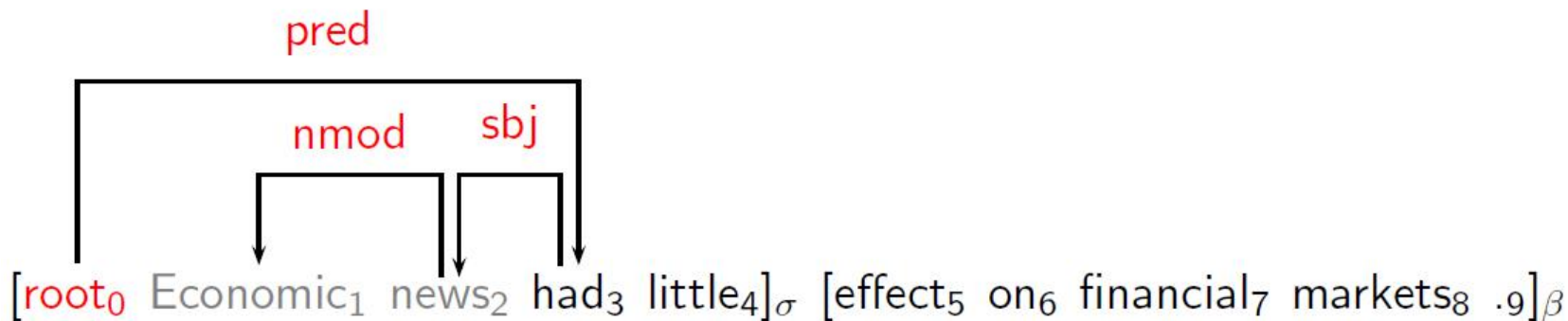
38



Right-Arc_{pred}

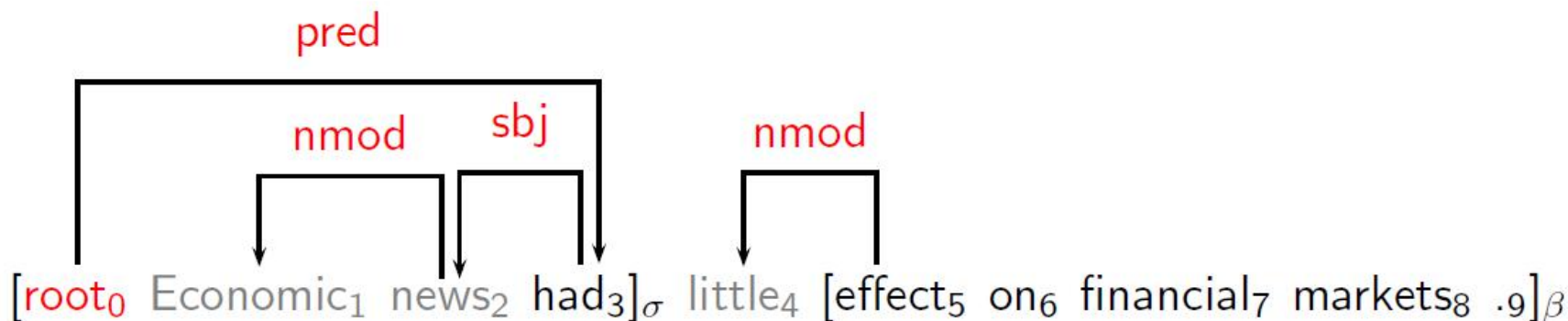
数据驱动-基于转移的方法

39



数据驱动-基于转移的方法

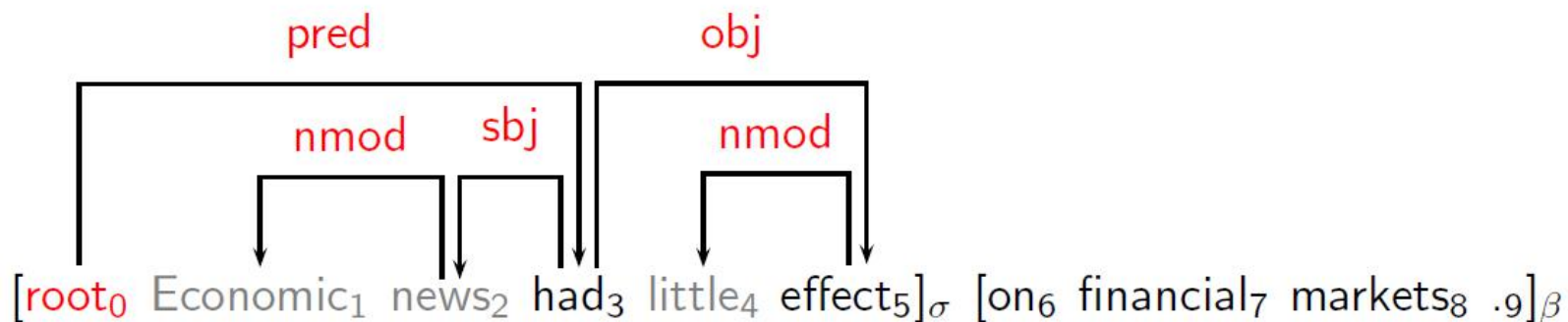
40



Left-Arc_{nmod}

数据驱动-基于转移的方法

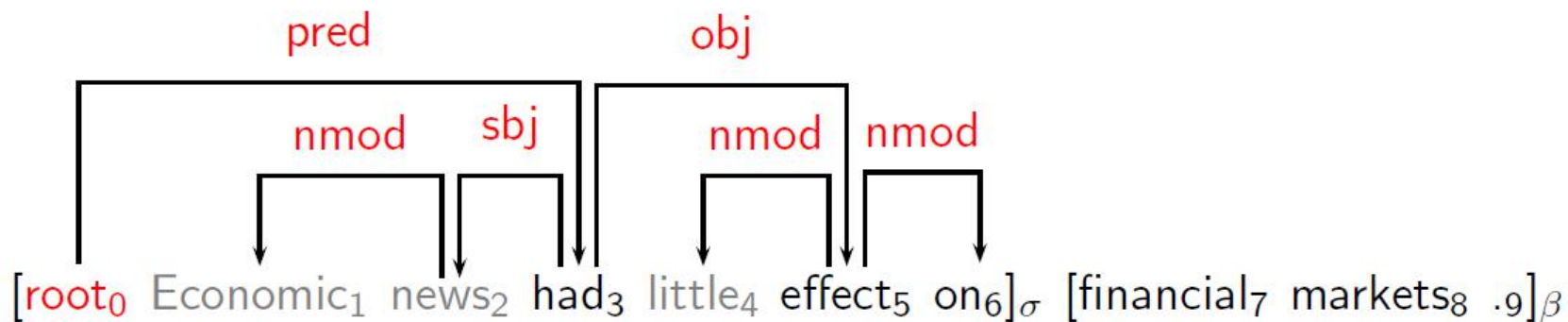
41



Right-Arc_{obj}

数据驱动-基于转移的方法

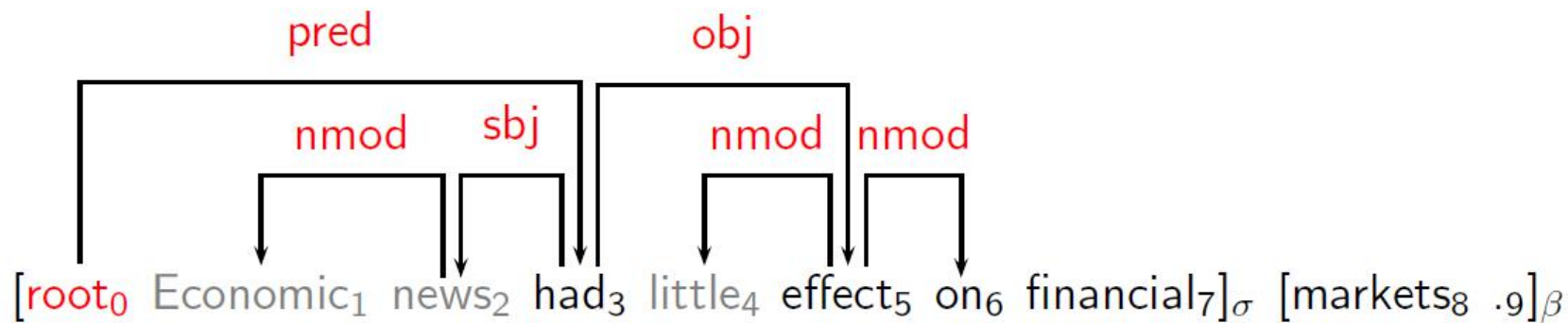
42



Right-Arc_{nmod}

数据驱动-基于转移的方法

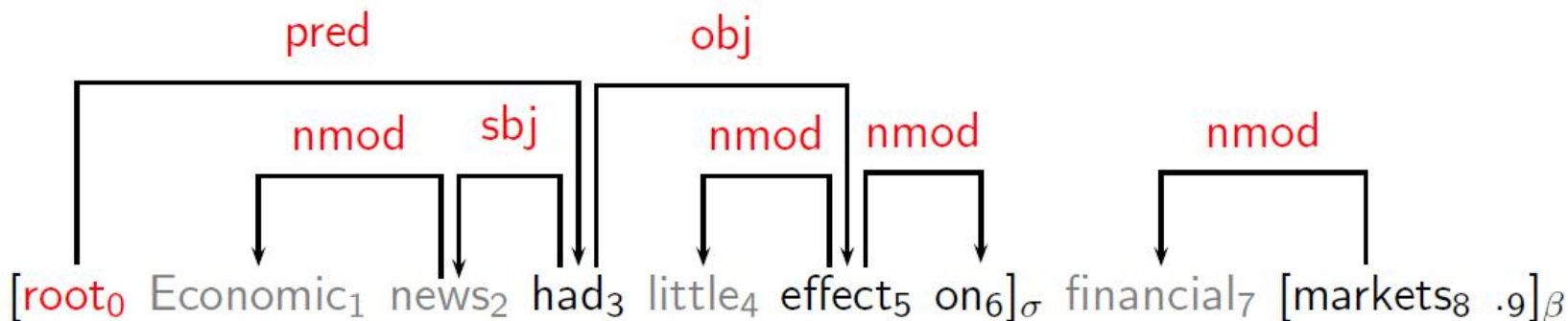
43



Shift

数据驱动-基于转移的方法

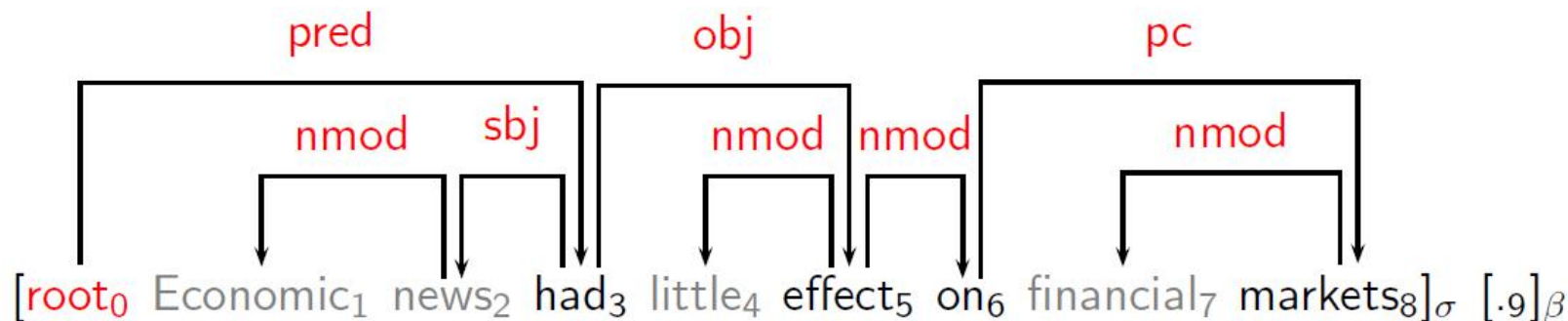
44



Left-Arc_{nmod}

数据驱动-基于转移的方法

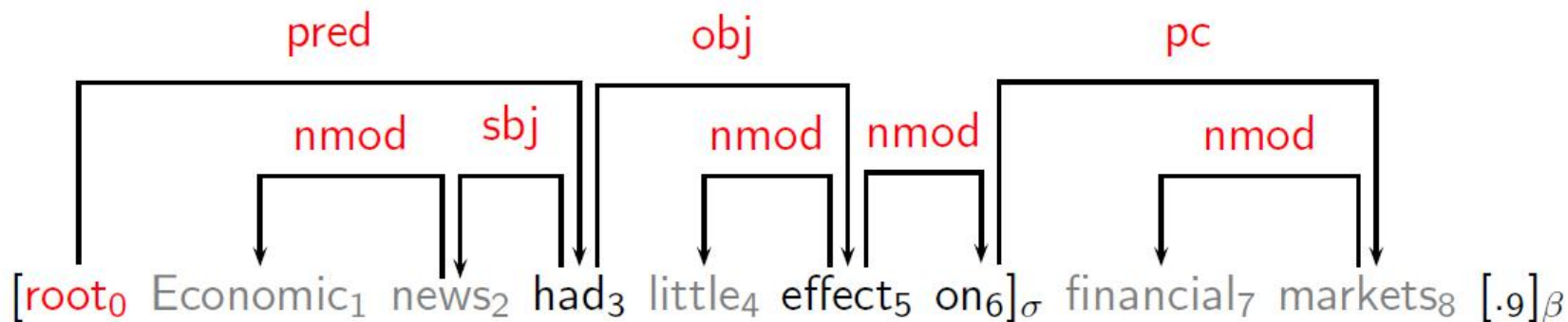
45



Right-Arc_{pc}

数据驱动-基于转移的方法

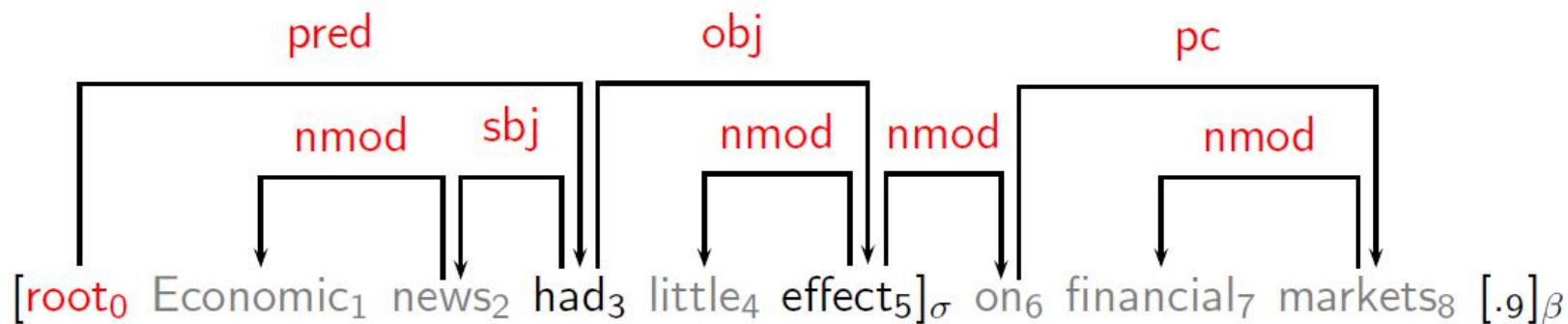
46



Reduce

数据驱动-基于转移的方法

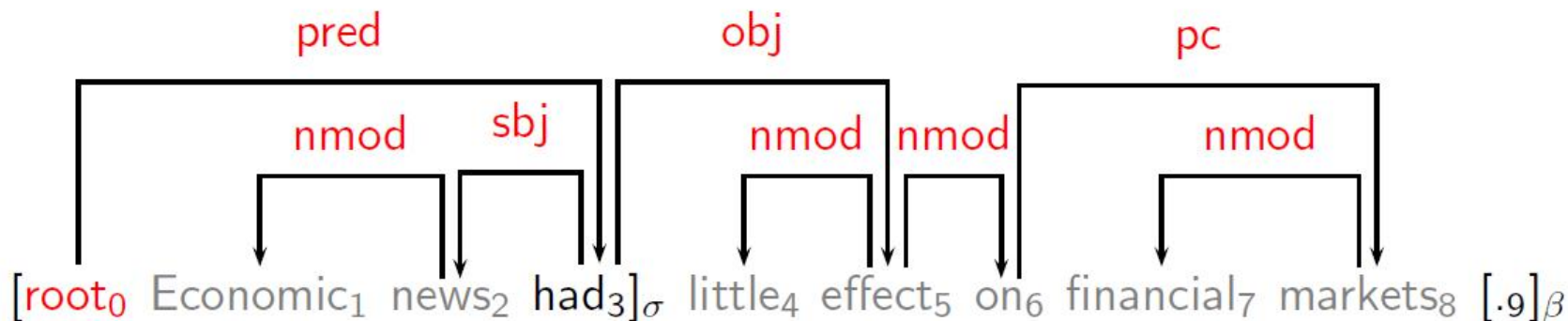
47



Reduce

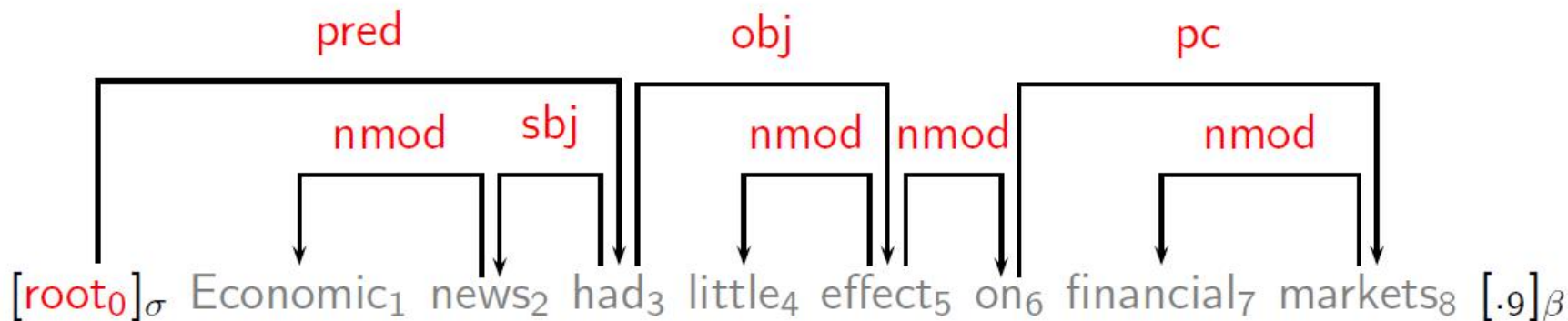
数据驱动-基于转移的方法

48



数据驱动-基于转移的方法

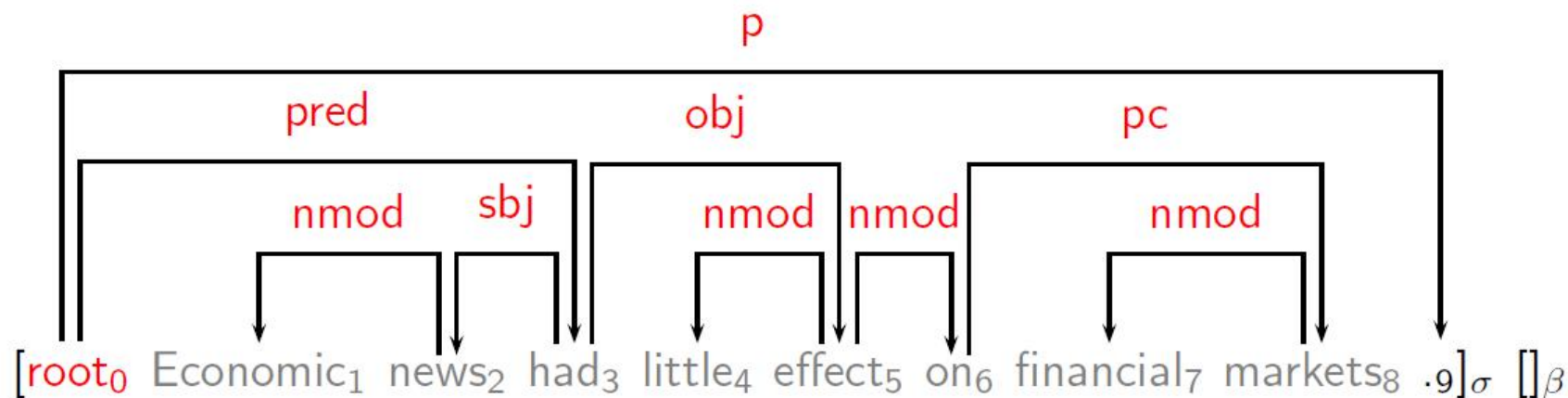
49



Reduce

数据驱动-基于转移的方法

50



Right-Arc_p

数据驱动-基于转移的方法

- (1) 依存句法结构描述
 - 预测结构变为预测动作序列
- (2) 分析算法设计与实现
 - 模型直接预测动作序列
- (3) 文法规则或参数学习
 - 和真实的动作序列做损失

3 短语结构与依存结构的关系

3 短语结构与依存结构的关系

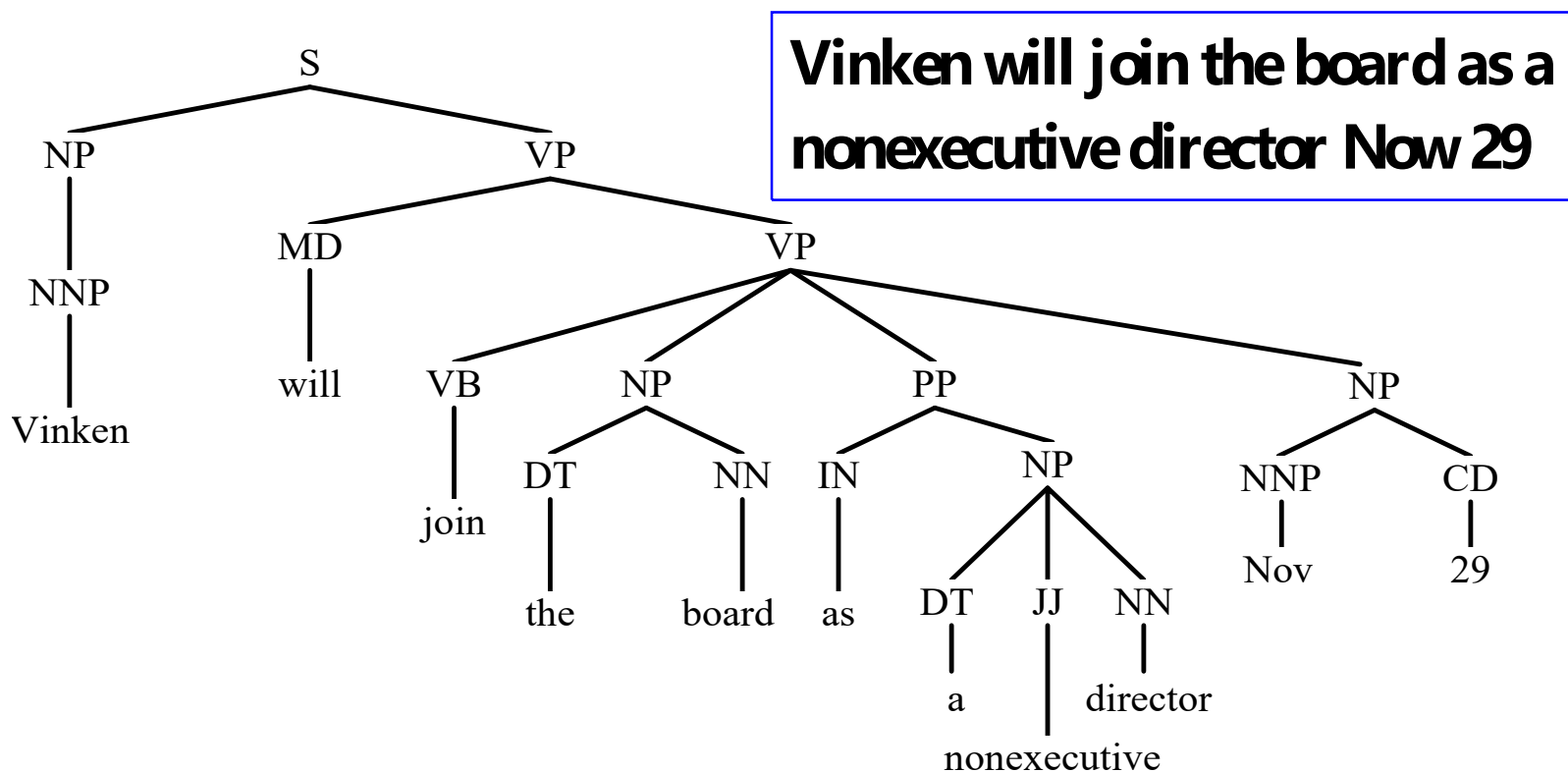
53

- 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，为句法树中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。

3 短语结构与依存结构的关系

54

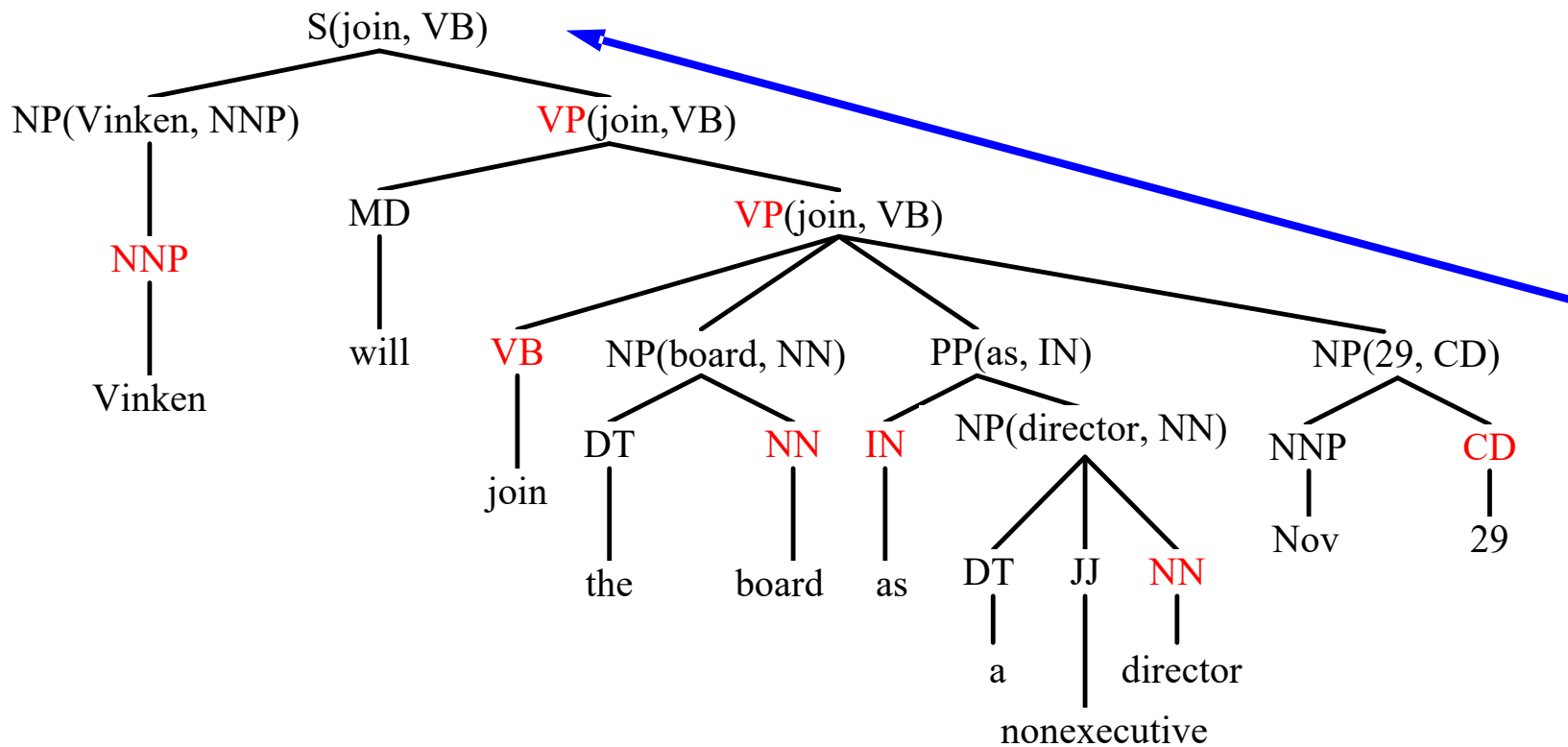
□ 例如：给定如下短语结构树



3 短语结构与依存结构的关系

55

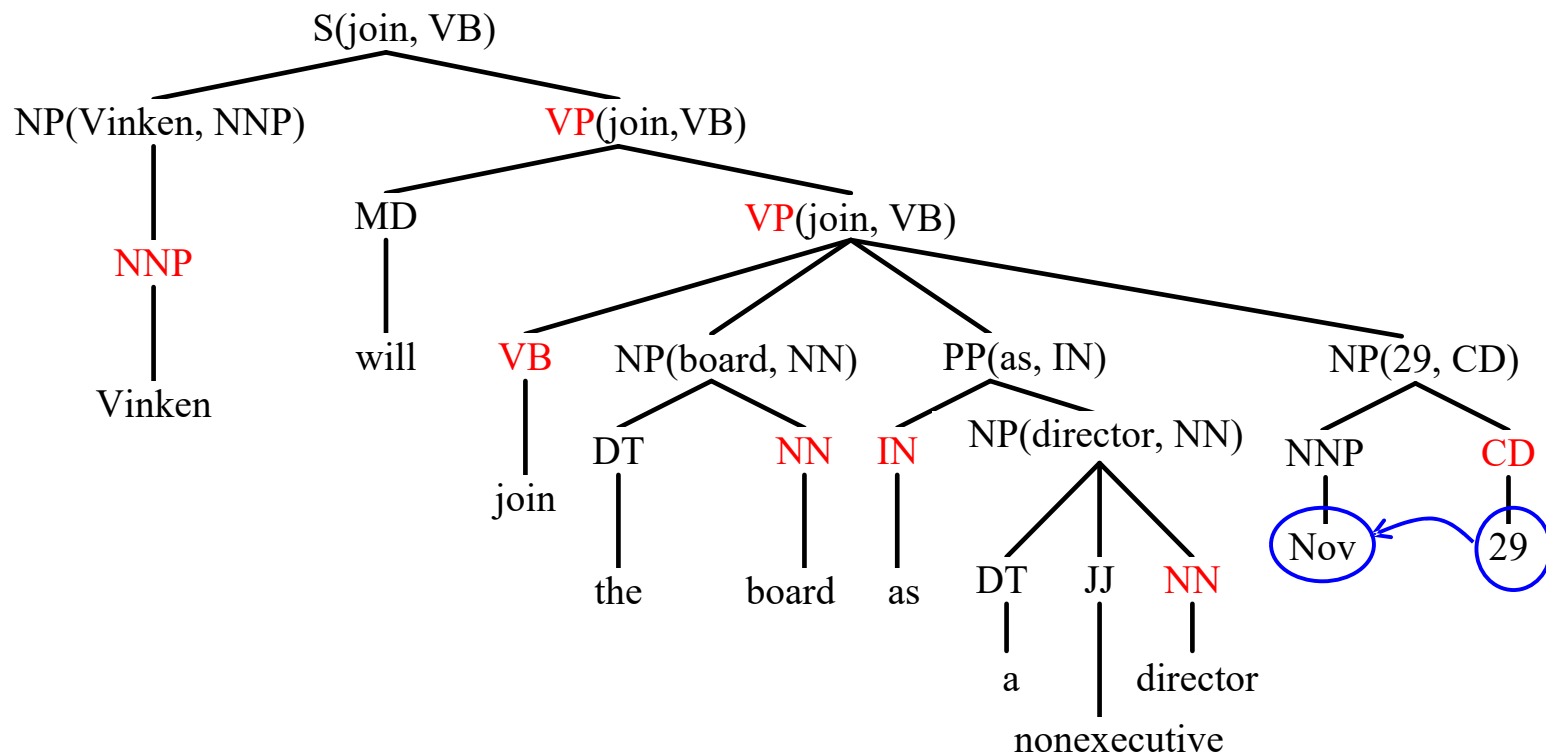
- 根据中心词表为每个节点选择中心子节点
(中心词通过自底向上传递得到)



3 短语结构与依存结构的关系

56

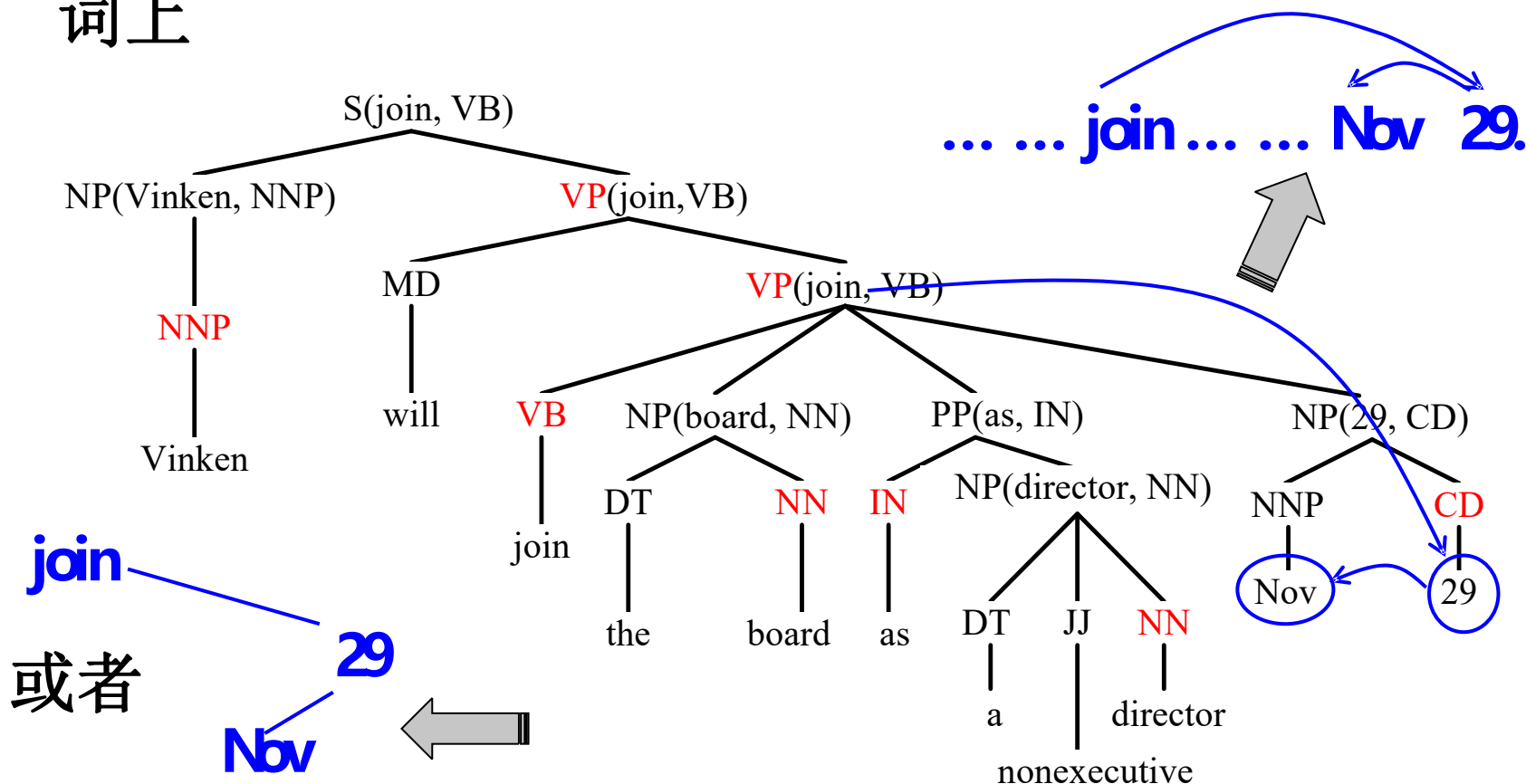
- 将非中心子节点的中心词依存到中心子节点的中心词上



3 短语结构与依存结构的关系

57

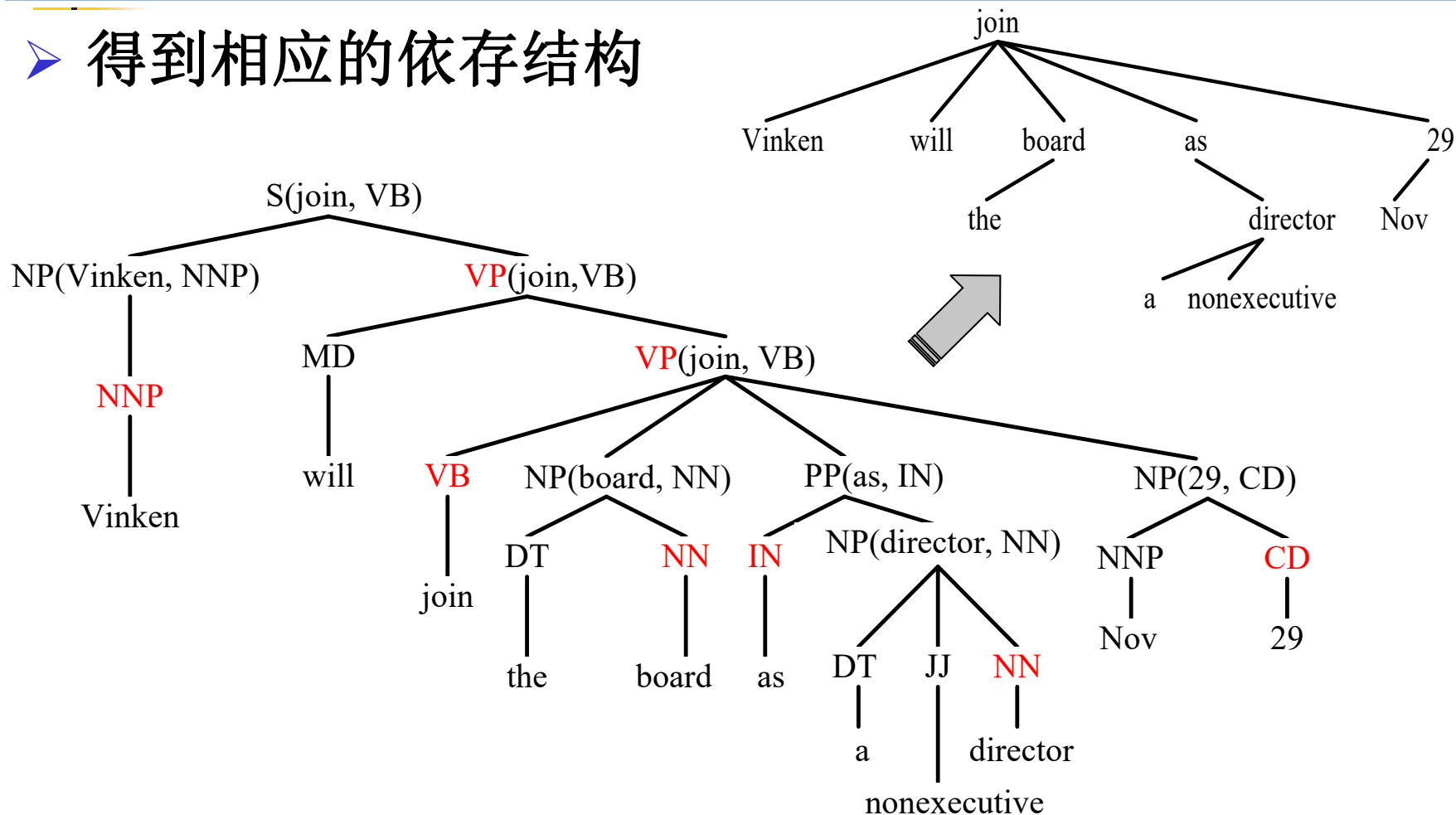
- 将非中心子节点的中心词依存到中心子节点的中心词上



3 短语结构与依存结构的关系

58

➤ 得到相应的依存结构



3 短语结构与依存结构的关系

59

- 利用高阶依存信息构造句法重排序模型
- 基本思想：
 - ▣ 在传统的基于PCFG的句法分析器中，要求的独立性假设过强，缺乏上下文词汇信息的帮助；
 - ▣ 词汇化方法LPCFG(Lexicalized PCFG)引入了中心词信息，改善了PCFG方法的性能。
 - ▣ 但是，中心词所携带的信息是有限的；
 - ▣ 可通过引入词汇依存信息，进一步改善句法分析器的性能。

3 短语结构与依存结构的关系

60

- 实现方法：利用判别式句法分析模型
- 句法树的评价得分Score可以通过下式计算：

$$Score(x, c) = \Phi(x, c) \cdot \bar{\alpha}$$

- 其中 x 为输入句子， c 为句法树；
 - $\Phi(x, y) \in \mathbb{R}^d$ 为句法树的特征向量；
 - $\bar{\alpha} \in \mathbb{R}^d$ 为特征权值向量。
- 句法分析的过程就是找到得分最高的句法树：

$$c^* = \arg \max_{c \in GEN(x)} Score(x, c)$$

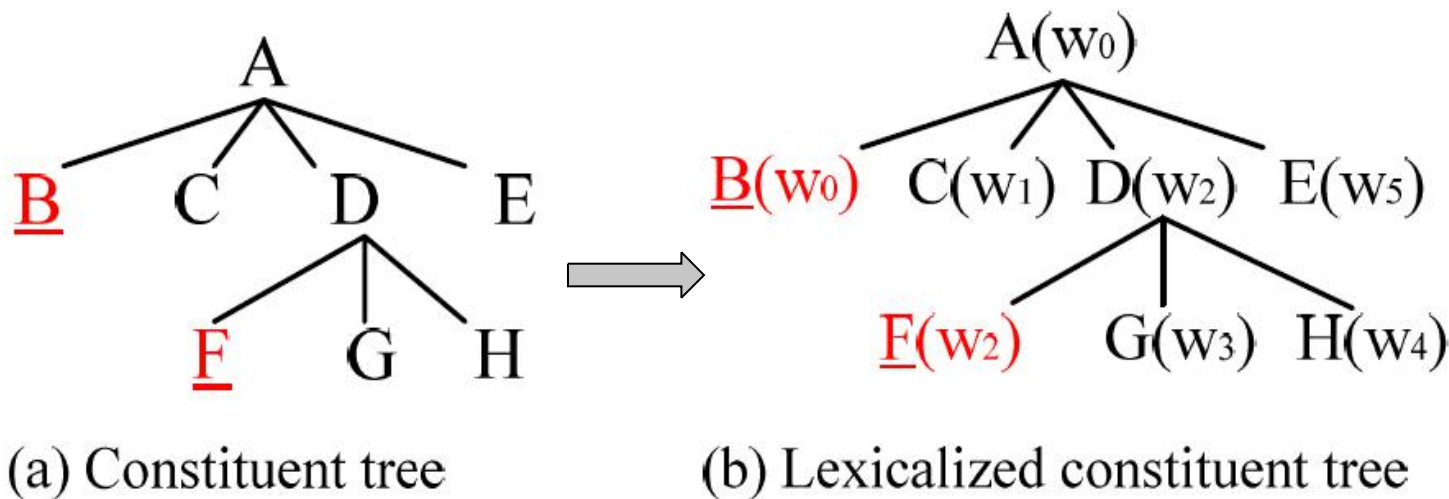
- 其中，函数 $GEN(x)$ 用于为句子 x 枚举出句法树候选。

3 短语结构与依存结构的关系

61

□ 将短语结构树转换为对应的依存分析树：

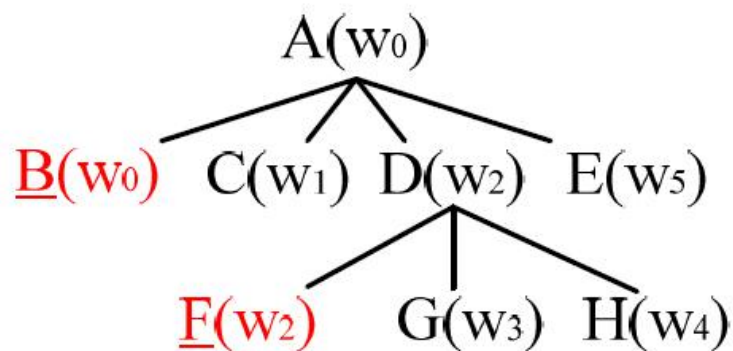
▣ 第一步：词汇化短语结构树



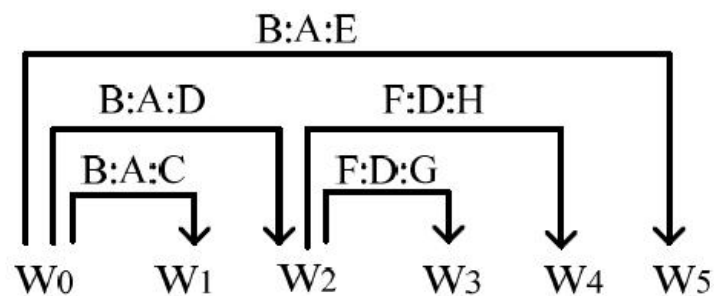
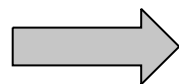
3 短语结构与依存结构的关系

62

- 将短语结构树转换为对应的依存分析树：
 - ▣ **第二步**：将词汇化的短语树转换为有标记的依存分析树



(b) Lexicalized constituent tree



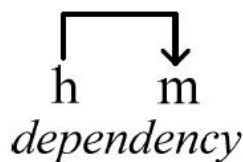
(c) Labeled dependency tree

3 短语结构与依存结构的关系

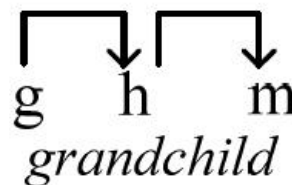
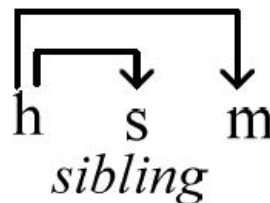
63

□ 各种词汇依存结构

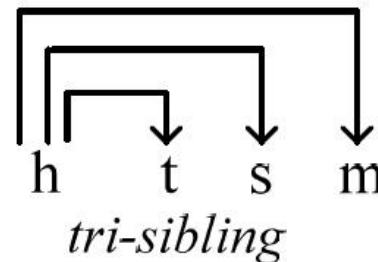
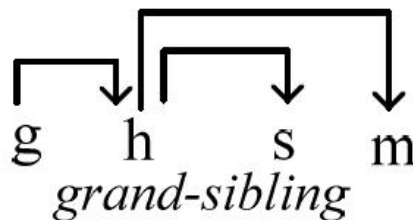
□ 一阶词汇依存结构



□ 二阶词汇依存结构



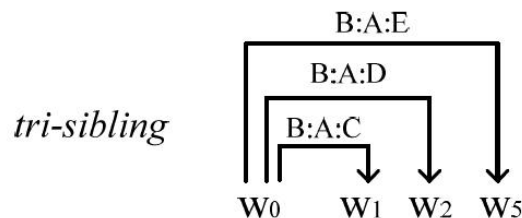
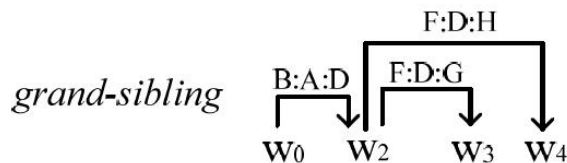
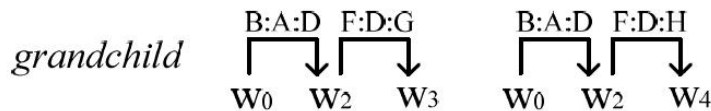
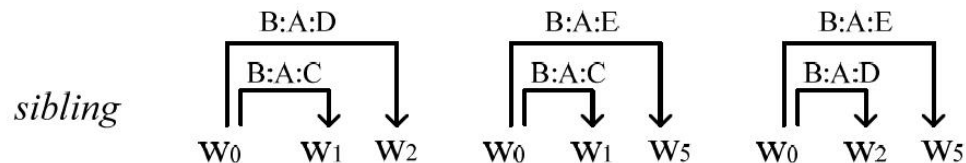
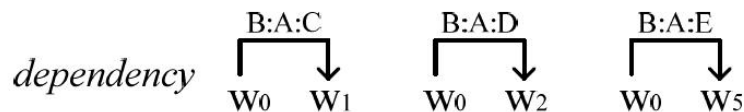
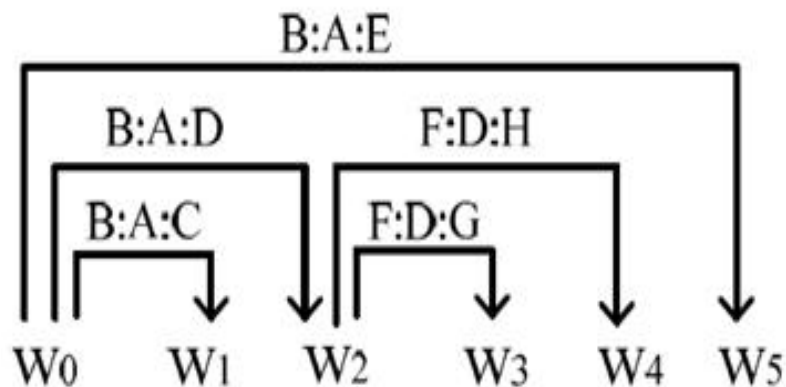
□ 三阶词汇依存结构



3 短语结构与依存结构的关系

64

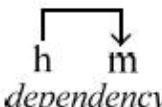
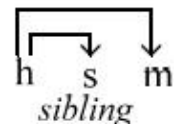
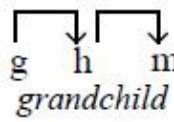
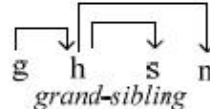
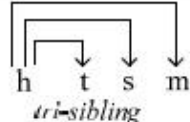
- 抽取词汇依存结构
- 如前图中节点A的各种词汇依存结构为：



3 短语结构与依存结构的关系

65

□ 定义词汇依存特征模板，构造特征向量

 dependency	Basic Uni-gram Features h , POS(h), N(h) h , POS(h) h , N(h) m , POS(m), N(m) m , POS(m) m , N(m)	 sibling	POS(h),N(h),POS(s),N(s),P(s),POS(m),N(m),P(m) POS(h),N(h),N(s),P(s),N(m),P(m) POS(h),N(h),POS(s),P(s),POS(m),P(m) POS(h),N(h),POS(s),N(s),POS(m),N(m) POS(h),POS(s),POS(m) N(h),N(s),N(m) N(h),P(s),P(m)
	Basic Bi-gram Features P(m) , h , POS(h), N(h), m), POS(m), N(m) h , POS(h), N(h), m , POS(m), N(m) P(m) ,POS(h), N(h), POS(m), N(m) P(m) , h , N(h), m , N(m) P(m) ,h , POS(h), m , POS(m) P(m) ,h , m P(m) , POS(h),POS(m) P(m) , N(h), N(m)	 grandchild	POS(g),N(g),POS(h),N(h),P(h),POS(m),N(m),P(m) POS(g),N(g),N(h),P(h),N(m),P(m) POS(g),N(g),POS(h),P(h),POS(m),P(m) POS(g),N(g),POS(h),N(h),POS(m),N(m) POS(g),POS(h),POS(m) N(g),N(h),N(m) N(g),P(h),P(m)
	Surrounding Word POS Features P(m), N(h), POS(h), N(m), POS(m), POS(h)+1, POS(m)-1 P(m), N(h), POS(h), N(m), POS(m), POS(h)-1, POS(m)-1 P(m), N(h), POS(h), N(m), POS(m), POS(h)+1, POS(m)+1 P(m), N(h), POS(h), N(m), POS(m), POS(h)-1, POS(m)+1	 grand-sibling	POS(g),POS(h),POS(s),POS(m) N(g),N(h),N(s),N(m) N(g),P(h),P(s),P(m)
		 tri-sibling	POS(h),POS(t),POS(s),POS(m) N(h),N(t),N(s),N(m) N(h),P(t),P(s),P(m)

4 汉英句法结构特点对比

4 汉英句法结构特点对比

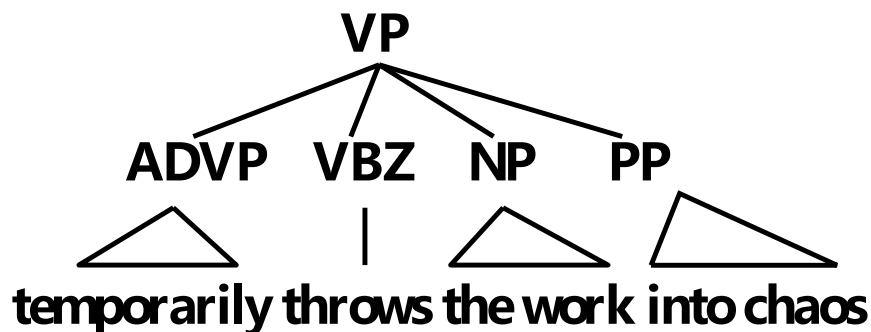
67

- 汉语比英语更少地使用功能词(function words), 且没有形态变化:
 - ▣ 汉语中不使用限定词(“这、这个、那个”等)的名词普遍存在, 复数标记(“们”等)有限并且很少出现。

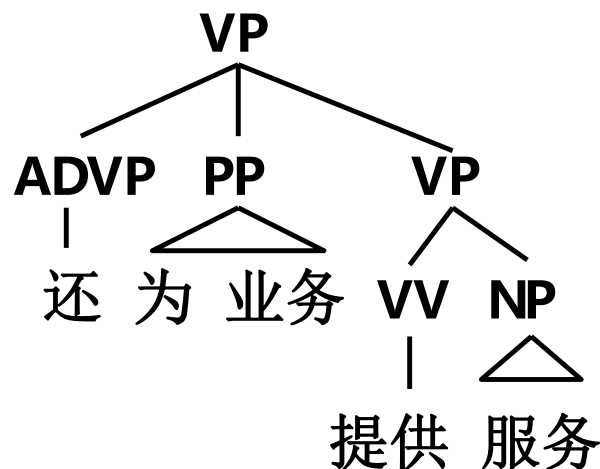
4 汉英句法结构特点对比

68

- 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：



(c)

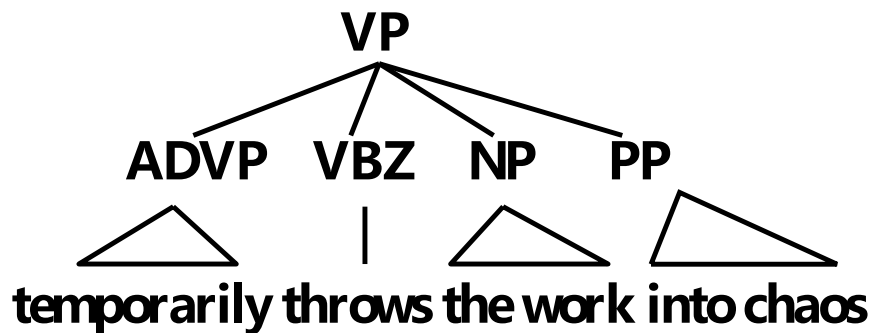


(d)

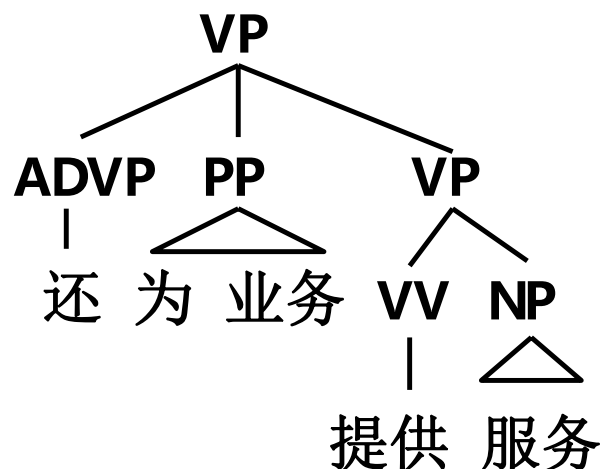
4 汉英句法结构特点对比

69

- 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，而在汉语句子里很少有这样歧义存在。



(c)



(d)

4 汉英句法结构特点对比

70

- 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(IP)结构还是仅仅是一个动词短语VP，如：

He thinks it is true. / 他认为X是对的。

4 汉英句法结构特点对比

71

□ 英语：“结构型”语言

- 一个完整的句法结构即表示一个完整的句子。
- 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：“表意型”语言

- 汉语句子通常受**语义**的牵引，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。
- 在汉语中存在一种**独特的长句构成方式**，就是一连串独立的简单句通过逗号或分号，连接成一个复杂的“句群”式的长句。

4 汉英句法结构特点对比

72

- 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。
- 因此，在很多情况下，它们之间的**分隔标记**仅仅是一个**逗号或者分号**。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。”

4 汉英句法结构特点对比

73

- 从中文资源联盟 (Chinese LDC) 发布的汉语树库(TCT 973)中随机地抽取出4431 个长度超过20个词的长句，其中，流水复句有1830个，占全部长句的41.3%[李幸, 2005]。

4 汉英句法结构特点对比

74

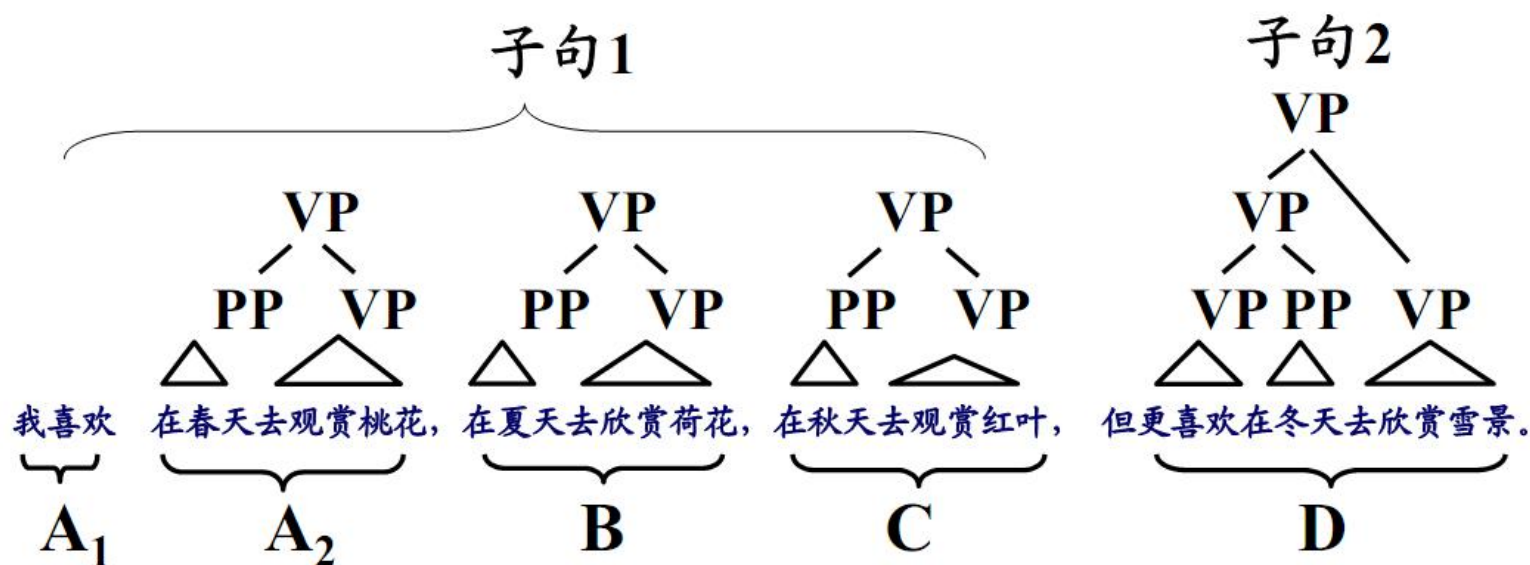
□ 汉语长句的层次化句法分析方法

- ▣ 对包含“分割”标点的长句进行分割；
- ▣ 对分割后的各个子句分别进行句法分析(即第一级分析)，分析得到的各个最大概率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；
- ▣ 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。

4 汉英句法结构特点对比

75

例句：我喜欢在春天去观赏桃花，在夏天去欣赏荷花，在秋天去观赏红叶，但更喜欢在冬天去欣赏雪景。



5 局部句法分析

和序列标注强相关，建议自行进一步学习

5 局部句法分析

77

- S. Abney (1991) 首先提出了浅层句法分析的概念。
- 浅层句法分析也称部分句法分析 (partial parsing) 或语块划分 (chunking)
 - ▣ 与完全句法分析不同
 - ▣ 只要求识别句子中某些结构相对简单的独立成分，如：非递归的名词短语、动词短语等。
- 浅层句法分析将句法分析任务分解为两个子任务：
 - ▣ ①语块的识别和分析；
 - ▣ ②语块之间的依附关系分析。

5 局部句法分析

78

- 语块：是介于字和句子之间的具有非递归特征的核心成分。
- 英语语块的定义包含三个层次：
 - ▣ 词 (words)
 - ▣ 非递归的名词短语(NPs)、动词词组(VGs)、副词短语(DPs)和介词短语(PPs)
 - ▣ 子句(clause)

5 局部句法分析

79

□ Base NP 定义

- ▣ 简单的、非嵌套的名词短语，不含其它的子短语。
- ▣ 特点：
 - 中心语：名词；
 - 不含其他的子项短语，**base NP**之间结构上是独立的。

5 局部句法分析

80

□ Base NP的形式化定义

- $\text{base NP} \rightarrow \text{base NP} + \text{base NP}$
- $\text{base NP} \rightarrow \text{base NP} + \text{名词} | \text{名动词}$
- $\text{base NP} \rightarrow \text{限定性定词} + \text{base NP} | \text{名词}$
- $\text{base NP} \rightarrow \text{限定性定词} + \text{名词} | \text{名动词}$
- $\text{限定性定词} \rightarrow \text{形容词} | \text{区别词} | \text{动词} | \text{名词} | \text{处所词} | \text{数量词} | \text{外文字串} | \text{数词和量词}$

5 局部句法分析

81

□ Base NP的识别

- ▣ 句子中的成分可以简单地分成base NP和非base NP两类
- ▣ 识别就成为一个分类问题

□ 两种表示方法

- ▣ 分隔法
- ▣ IOB标注方法

5 局部句法分析

82

- 例1: [Pierre Vinken],[61 years] old, will join [the board] as [a non-executive director] on [Nov. 29].
- 例2: When [it] is [time] for [their biannual powwow], [the nation]'s [manufacturing titans] typically jet off to [the sunny confines] of [resort towns] like [Boca Raton and Hot Springs].
- 例3: 一个于[半个世纪]之后重新聚集在“[西南联大]”[旗帜]下的[奉献活动]开始了!

5 局部句法分析

83

□ IOB标注方法中，

- 字母 ‘B’ (Begin)表示当前词语位于base NP的开端，
- 字母 ‘I’(In)表示当前词语在base NP内 (非短语首词语)，
- 字母 ‘O’ (Out)表示词语位于base NP 之外。例如：

例4：外商/B 投资/I 成为/O 中国/B 外贸/I 重要/B 增长/I。
/O

□ 与IOB方法类似的标注方法还有：

- IOE (In, Out, End)表示方法，
- Start/End表示方法(用5个标志, O, B, E, I, S) 等。

5 局部句法分析

84

□ Base NP识别方法

- ▣ 基于SVM的识别方法
- ▣ 基于WINNOWN的识别方法
- ▣ 基于CRF的识别方法

5 局部句法分析

85

□ 基于SVM的识别方法

- ▣ SVM: 二值分类

- ▣ base NP: 多值分类

- ▣ 转化策略:

 - 配对策略

 - 一比其余策略

5 局部句法分析

86

□ 基于SVM的识别方法

- ▣ T. Kudo等(2003)

- ▣ YamCha系统

- ▣ 三类特征

- 词: $w_{i-1}w_{i-1}w_iw_{i+1}w_{i+2}$

- 词性: $t_{i-1}t_{i-1}t_it_{i+1}t_{i+2}$

- base NP标志: $c_{i-2}c_{i-1}$

其中, w_i 为句子中位置*i*处的词, t_i 为词的词性, c_i 为要识别的第*i*个词的base NP标记。

5 局部句法分析

87

□ YamCha 系统识别base NP过程

	COL: 0	COL: 1	TAG	
POS:-4	He	PRP	B-NP	
POS:-3	reckons	VBZ	B-VP	
POS:-2	the	DT	B-NP	Feature Sets
POS:-1	current	JJ	I-NP	
POS: 0	deficit	NN	I-NP	Eestimated TAG
POS:+1	will	MD	B-VP	
POS:+2	narrow	VB	I-NP	
POS:+3	to	TO	B-PP	

5 局部句法分析

88

□ 用于base NP识别语料资源

- 英文：CoNLL-2000 (Conference on Computational Natural Language Learning)提供的《华尔街日报》语料
 - 训练语料：15—18章，211,727个词
 - <http://www.cnts.ua.ac.be/conll2000/chunking/train.txt.gz>
 - 测试语料：第20章，47,377个词
 - <http://www.cnts.ua.ac.be/conll2000/chunking/test.txt.gz>
 - 工具Chunklink 用于语料格式转换。
 - <http://ilk.kub.nl/~sabine/chunklink/>
- 汉语：宾州 LDC 语料库，中文树库

5 局部句法分析

89

- 根据有关实验，最好的汉语base NP识别正确率比英语base NP识别的正确率相差约5%左右。主要原因：
 - ▣ 英语中大多数base NP有限定词，如形容词、冠词等，而汉语没有；
 - ▣ 带多个名词修饰语的汉语 base NP 识别不完整，如：“高新技术”，“高”可能无法被识别；
 - ▣ 连续名词序列组成的 base NP识别困难，如：中国/NR 红十字会/NR 名誉/NN
 - ▣ 汉语词性和语义更多样化。

小结

90

- 依存句法分析
- 汉语句法结构特点对比
- 局部句法分析