第 13 章　篇 章 分 析

中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

基本概念

★ 什么是篇章?

篇章：由一个以上的句子（sentence）或语段（utterance）构成的**有组织、有意义**的自然语言文本整体。一篇文章、一段会话等都可以看成篇章。构成篇章的句子（或语段）彼此之间在形式上相互衔接，在意义上前后连贯。

例1：小明学习刻苦，成绩每年进步，考上理想大学。√

例2：花是红的，人工智能飞速发展，今天傍晚有雨。×

例3：考上理想大学，成绩每年进步，小明学习刻苦。×

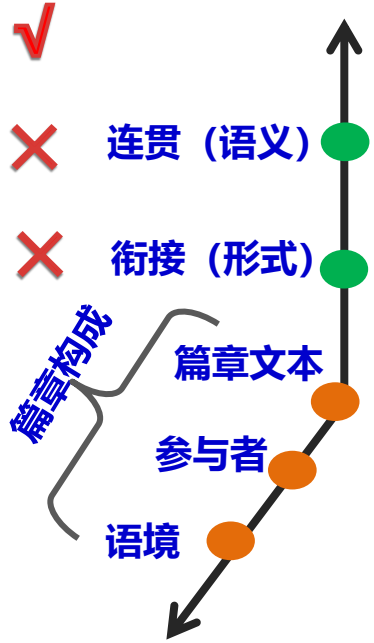
例4：对话

S1：电话铃响了。

S2：我正在看书。

S1：好的。

语境、意图 √



基本概念

✧ 篇章研究内容

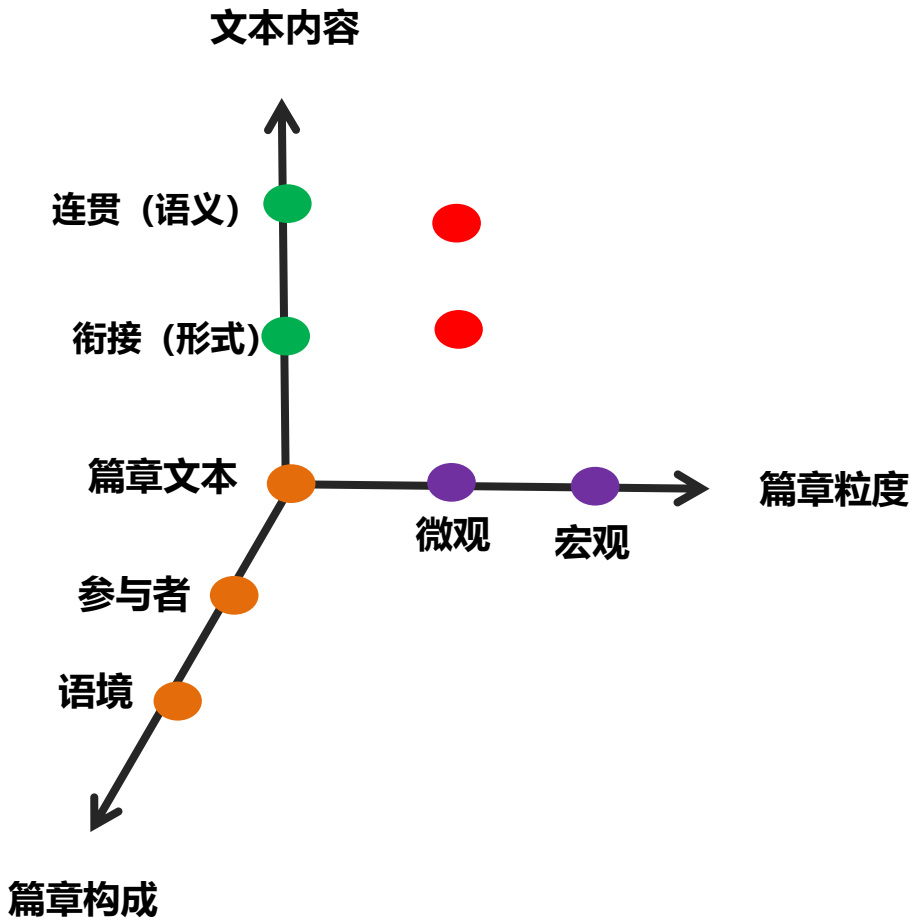
篇章的7个基本特征

篇章文本	**衔接性**（cohesion） **连贯性**（coherence）
参与者（产生者和接受者）	**意图性**（intentionality） **可接受性**（acceptability）
语境	**信息性**（informativity） **情景性**（situationlity） **跨篇章性**（intertextuality）

（de Beaugrande 和 Dressler，1981 ）

基本概念

✧ **篇章分析内容：**

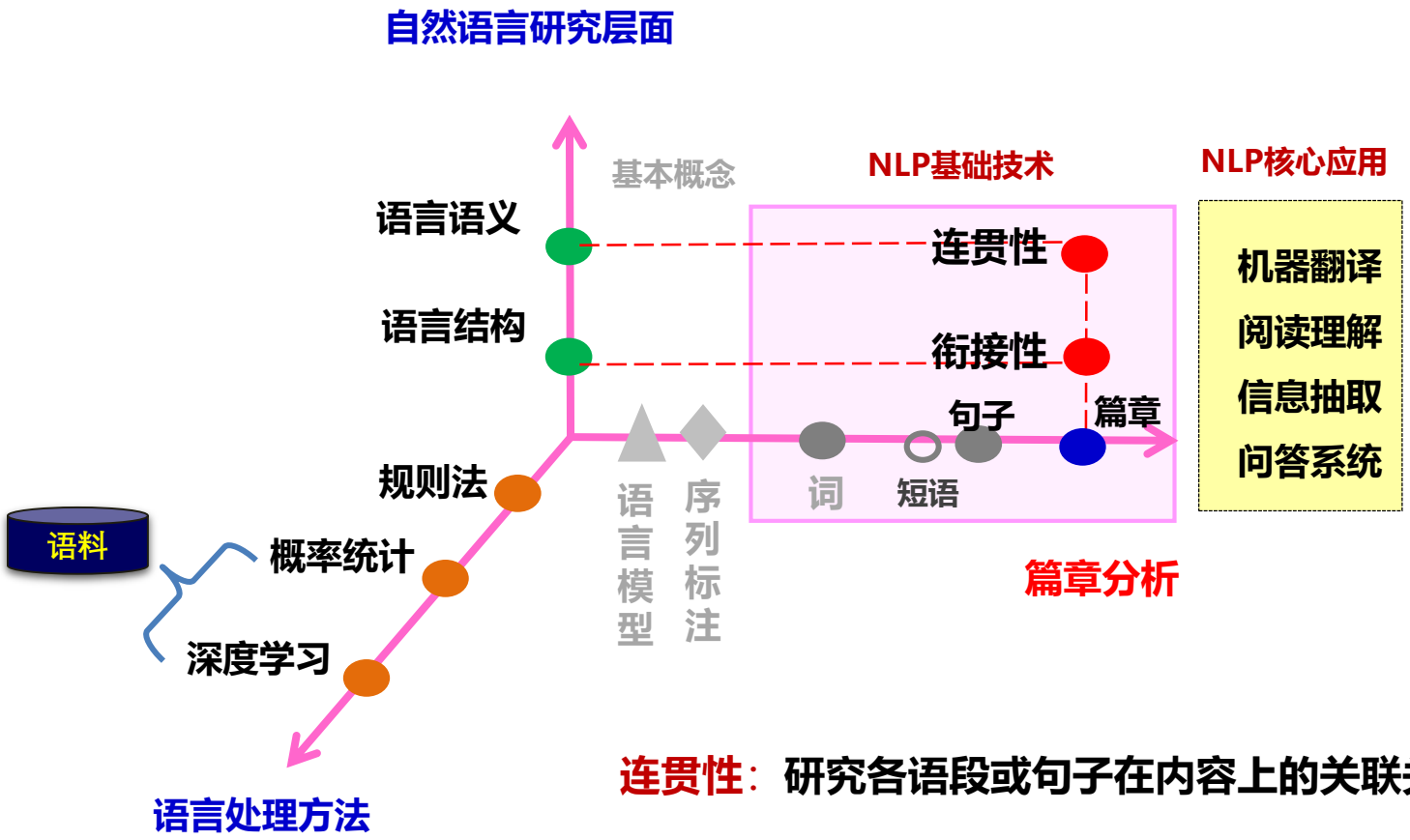


诸多篇章理论

浅层衔接理论，Hobbs模型、修辞结构理论、宾州篇章树库理论、意图结构理论、信息结构理论、主位述位理论、 D-LTAG理论、句群理论、中心理论、语言行为理论、复句理论和基于连接依存树的篇章结构理论；篇章模式、超主位理论和篇章宏观结构理论等.

自然语言处理课程内容及安排

✧ **课程内容：**



连贯性：研究各语段或句子在内容上的关联关系

衔接性：研究各语段或句子在形式上的关联关系

内 容 提 要

13.1　篇章连贯性分析

13.2　篇章衔接性分析

篇章连贯性概述

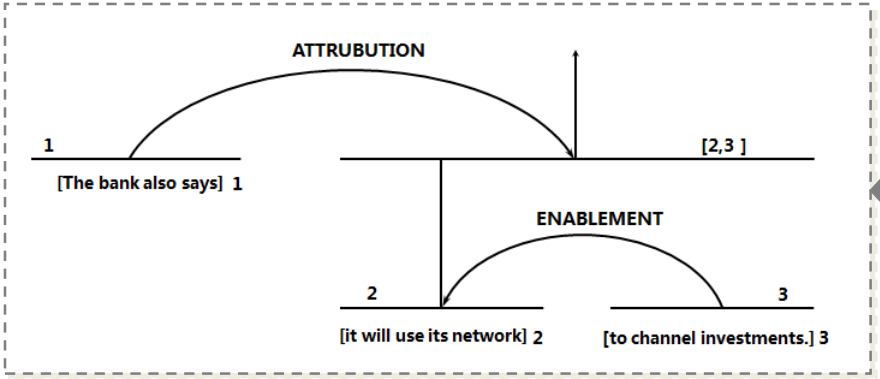
基于修辞结构理论篇章连贯性分析任务

输入(篇章): The bank also says it will use its network to channel investments

实现技术

- 概率统计
- 深度学习

输出:



修辞结构理论

篇章分析树

内 容 提 要

13.1　篇章连贯性分析

13.1.1 修辞结构理论 RST

13.1.2 基于RST的篇章分析法

13.2　篇章衔接性分析

13.1.1 修辞结构理论 RST

修辞结构理论 (Rhetorical Structure Theory, RST, Mann &Thompson ,1988**)**

- **RST理论关于语篇结构的基本设想概括起来有以下11点:**

1. **语篇由各具重要功能的部分构成**: 较小的部分按一定关系模式组成更大的部分,直至生成语篇。
2. 一段话语或文字要被确认为语篇,其**各个部分必须有机地结合以形成整体性和连贯性。**
3. **整体性与连贯性来自语篇的内在功能。**一个语篇之所以会产生整体性与连贯性的效果,是因为它的每一个组成部分都直接或间接地服务于语篇作者的同一中心目的。

13.1.1 修辞结构理论 RST

4 . 语篇的构成方式如下: 两个基本的部分组成一个较大的部分,这个较大的部分再与另一部分组成更大的部分,直至形成语篇。因此,语篇结构是层级结构,不是线性结构。

5. RST区别三种语篇结构,即类型结构(语篇的题材或类型特征所决定)、句法结构和关系结构。它描写关系结构及其同前两者的相互影响,而不对类型结构和句法结构作具体的分析和描写。

6. 关系结构内部具有同一性:从小句连接到语篇本身的所有层级结构共用一套相同的关系模式 。

7. 关系结构是多语句语篇的主要结构: 一套为数不多的递归性关系模式将基础的部分两两连接成更大的部分,直至形成语篇。

13.1.1 修辞结构理论 RST

8. **不对称性在关系结构中占主导地位。**英语中最常用的一类结构关系是RST中称为“Nucleus—Satellite”的不对称关系,即核心—辅助关系。

9. **语篇的结构关系是功能关系,不是形式关系。**其共同特点是可用不同的效果类型进行描写。描写角度可包括语篇作者的中心目的,作者对读者情况的假定等。

10. **语篇关系是语篇的深层结构关系。**这些抽象的关系由表层的语句来实现,但严格地讲,语篇关系并不存在于表层的语句或段落之间。

11. **语篇关系的种类和数量原则上是无限的**,以前未出现过的关系类型也可能在新的语篇关系中出现。但在实际应用中,绝大多数语篇均由一小部分复现率极高的常用关系摸式构成,换言之,用为数有限的关系就可分析一种语言的大部分语篇。如用20多种关系就可描写绝大多数英语语篇。

13.1.1 修辞结构理论 RST

RST术语的定义:

篇位(EDUs, Unit)—实现结构段的表层语篇单位, 一般为小句。

结构段(Span)—语篇结构中任何具有RST结构的、功能完整的片段。

核心结构段（Nucleus）— 在修辞关系中，一般满足一定关系的两个独立语段在其相互关系中所处的地位是不同，处于重要地位的语段称之为核心语段，一般用N 代表核心语段。 辅助结构段（Satellite）—在修辞关系中，处于辅助地位的语段称之为辅助语段，一般用 S 代表辅助语段。

13.1.1 修辞结构理论 RST

RST理论中最关键的两个成份: 关系和结构段

关系定义(Relation Definition) —确定两个结构段之间关系的依据和标准

具体包括两个方面:

1) **限制条件**(Constraints): 包括核心结构段限制条件、辅助结构段限制条件及这两种结构段联合限制条件

2) **效果**(Effect): 对作者使用某一关系所希望达到的效果及效果位置(Locus of Effect〕的说明。

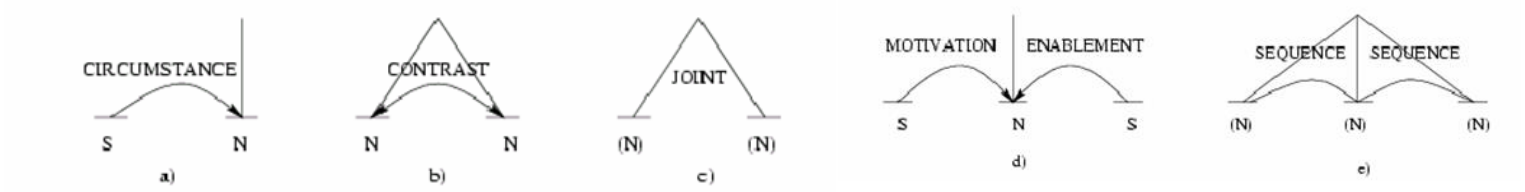
语篇结构(Text Structure) —由小到大的结构段关系系统。

13.1.1 修辞结构理论 RST

RST关系集:

经典的修辞关系集	
核心性	关系名称
单核心关系	证据/Evidence
	证明/Justification
	对照/Antithesis
	让步/Concession
	环境/Circumstance
	解答/Solution
	详述/Elaboration
	背景/Background
	使能/Enablement
	动机/Motivation
	目的/Purpose
	意愿性原因/Volitional cause
	非意愿性原因/Non-volitional cause
	意愿性结果/Volitional result

经典的修辞关系集	
核心性	关系名称
单核心关系	非意愿性原因/Non-volitional cause
	意愿性结果/Volitional result
	非意愿性结果/Non-volitional result
	条件/Condition
	否则/Otherwise
	解释/Interpretation
	评估/Evaluation
	重述/Restatement
	总结/Conclusion
多核心关系	序列/ Sequence
	对立/Contrast
	罗列/List
	联接/Joint



13.1.1 修辞结构理论 RST

关系内容定义

证据关系定义

核心结构段限制条件:读者有可能认为核心结构段可信度不足

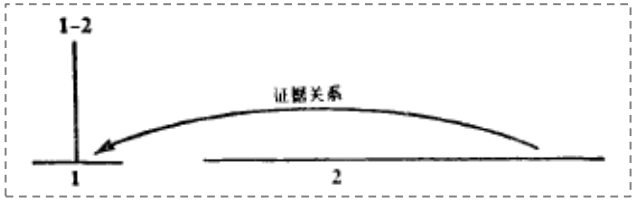
辅助结构段限制条件:读者认为辅助结构段是可信的

核心一辅助结构段联合限制条件:对辅助结构段的理解导致对核心结构段相信程度的增加

效果:读者对核心结构段的相信程度增加

效果位置:核心结构段

例：1.They are having a party again next door .
2. I couldn’t fingd a Parking Space.



RST关系图解

13.1.1 修辞结构理论 RST

环境关系定义

核心结构段限制条件: 无

辅助结构段限制条件: 辅助结构段提供一个(已实现的)情况

(辅助结构段为核心结构段的理解提供时间、地点等框架。)

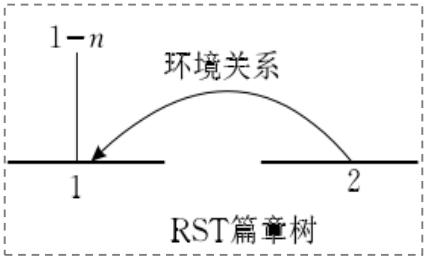
核心—辅助结构段联合限制条件: 读者应在辅助结构段设定的主题框架内理解核心结构段的内容

效果: 读者意识到应在辅助结构段提供的主题框架内理解核心结构段的内容

效果位置: 核心结构段和辅助结构段

例: 1.Probably the most extreme case of Visitors Fever I have evel witnessed was a few summer ago.

2. When I visited relatives in the Midwest.



RST篇章树

13.1.1 修辞结构理论 RST

序列关系定义

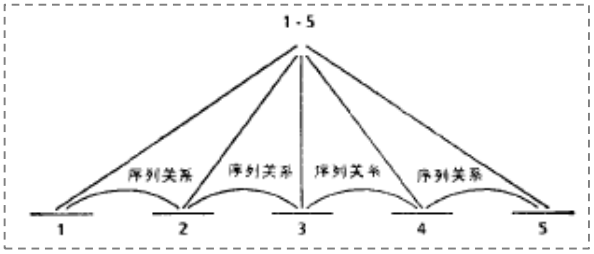
核心结构段限制条件: 多核心

多核心联合限制条件: 多核心所述情况呈序列分布

效果: 读者意识到多核心间的序列关系

效果位置: 多核心

例: 1.Peel oranges
2.And slice crosswise.
3 Arrange in a bowl
4.And sprinkle with rum and coeonut.
5.Chill until ready to serve.



RST关系图解

13.1.1 修辞结构理论 RST

例1：用 RST 理论分析如下篇章结构

(0) Farm prices in October edged up 0.7% from September，

(1) as raw milk prices continued to rise，

(2) the Agriculture Department said.

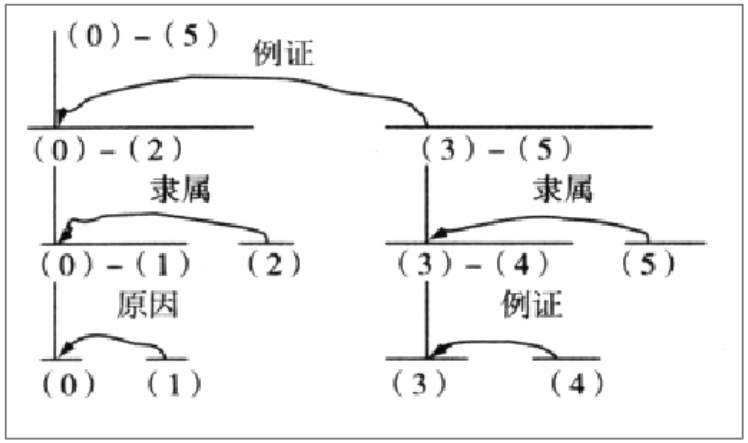
(3) Milk sold to the nation's dairy plants and dealers averaged $14. 50 for each hundred pounds，

(4) up 50 percent from September and up $ 1． 50 from October 1988，

(5) the department said.

解：

6 个基本语篇单位:



篇章结构树-RST关系图解

内 容 提 要

13.1　篇章连贯性分析

13.1.1 修辞结构理论 RST

13.1.2 基于RST的篇章分析法

13.2　篇章衔接性分析

13.1.2 基于RST的篇章分析法

基于RST的篇章结构分析主要包括两个子任务：

基本篇章单位EDUs的划分 和 篇章结构的生成.

1.篇位切分:

将整个语篇切分成若干篇位(EDUs)。篇位根据需要可大可小,

Mann&ThomsPon 以小句(Clause)作为基本篇位.

2.确定结构段

判定两个结构段之间的关系（可以是自上而下,也可以是自下而上或两者兼用）,在分析者在每一步都要考虑关系定义运用得是否合情合理。

13.1.2 基于RST的篇章分析法

例 1：句子级（内部）篇章分析

如：The bank also says it will use its network to channel investments

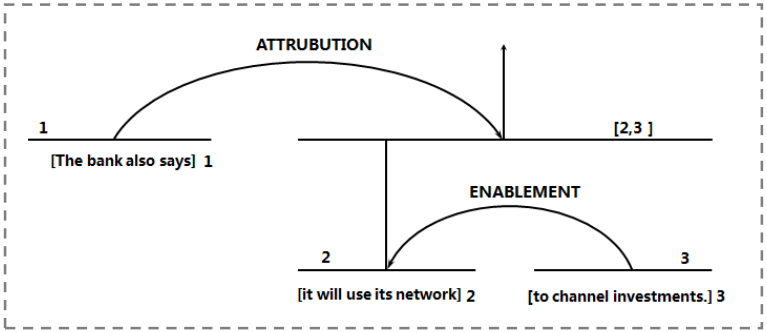
分析结果示例：

1. 篇位切分为:

< The bank also says> < it will use its network>< to channel investments>

2.确定结构段

自下而上判定两个结构段之间的关系,形成篇章分析树



篇章分析树

语篇核心

从RST结构的最顶端开始,沿竖线不间断地一直向下,到没有竖线接下去的篇位,便是语篇的核心篇位

Radu Soricut and Daniel Marcu，Sentence Level Discourse Parsing using Syntacticand Lexical Information

13.1.2 基于RST的篇章分析法

具体分析步骤：

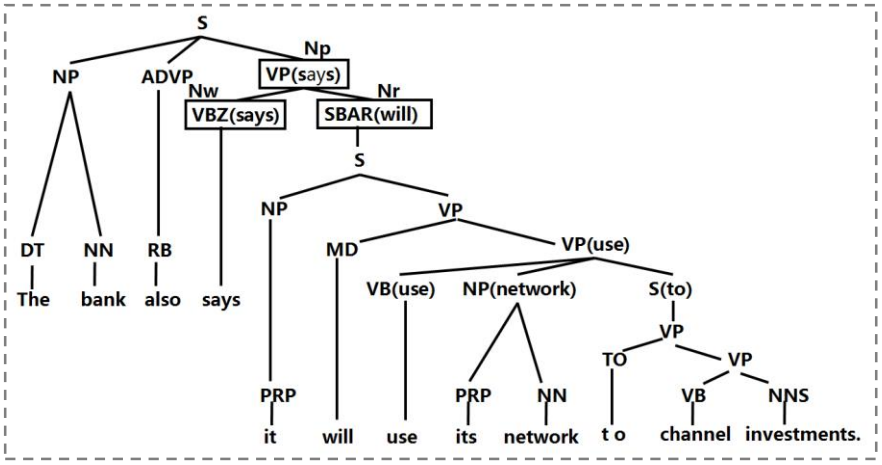
The bank also says it will use its network to channel investments

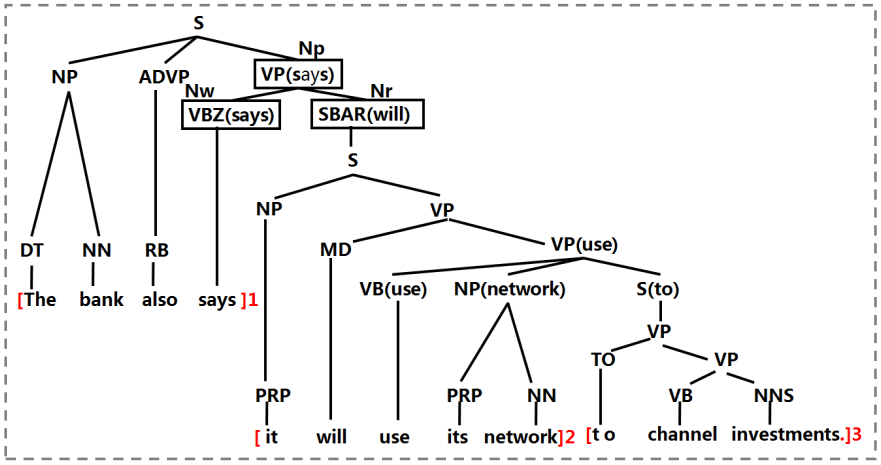
1. 篇位切分为目标:

输入： 句法树
(lexicalized syntactic trees.)

输出：在输入的句法树的篇位边界词后加边界标识符

< The bank also says>
< it will use its network>
< to channel investments>





13.1.2 基于RST的篇章分析法

方法:

（1）. 定义语段模型 计算句法树中句子每个词后 插入边界符的概率

（2）. 在句法树中每个 插入概率 大于 0.5 的词后插入边界符

A. 语段模型定义

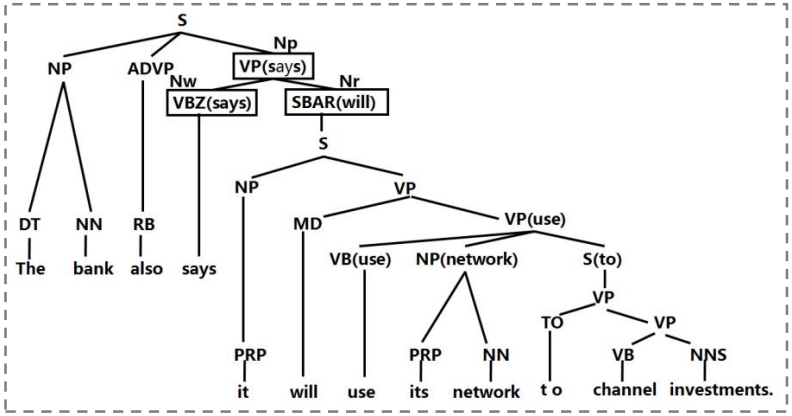
给定句子 S= $W_1W_1 \ldots\ldots W_i \ldots\ldots W_n$

t 为 s 的句法树

P(bi |,Wi, ,t) 表示Wi 是边界的概率 。

其中，bi ∈{ 边界，非边界}

P(边界 |,Wn, ,t)=1 句尾是语段边界



13.1.2 基于RST的篇章分析法

对于每个词 w，Nw 定义为在词汇化结构树中最上结点为w

且有右兄弟结点的结点； Np为Nw的父结点； Nr为Nw的右兄弟结点；

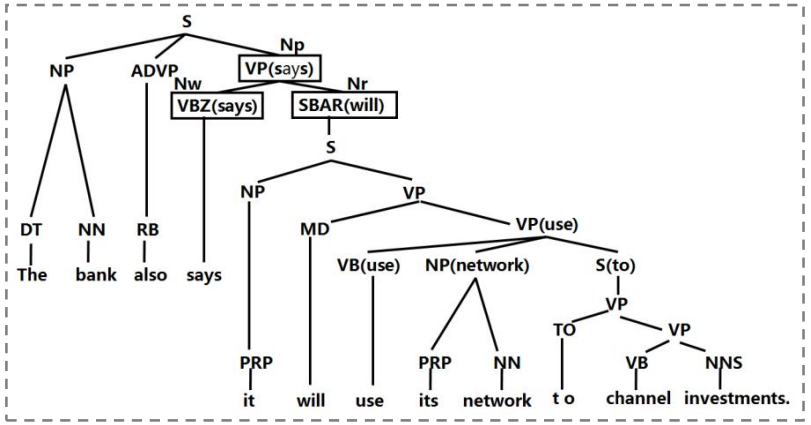
如： W=says $N_w = \texttt{VBZ(says)}$

$N_p = \texttt{VP(says)}$ $N_r = \texttt{SBAR(will)}$

$$P(b|w,t) \simeq \frac{Cnt(N_p \rightarrow \ldots N_w \uparrow N_r \ldots)}{Cnt(N_p \rightarrow \ldots N_w N_r \ldots)}$$

其中： $N_p \rightarrow \ldots N_w N_r \ldots$

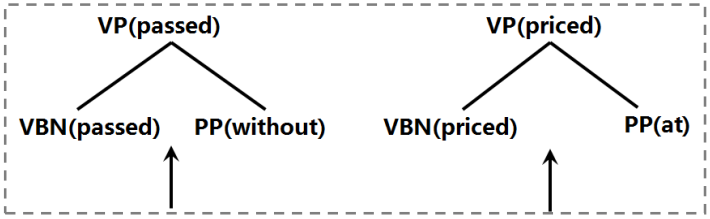
为词汇化规则



用语段模型计算每个词的边界概率，

在概率>0.5的词后插入边界符

需要词汇化规则理由：



13.1.2 基于RST的篇章分析法

B. 语段模型参数学习

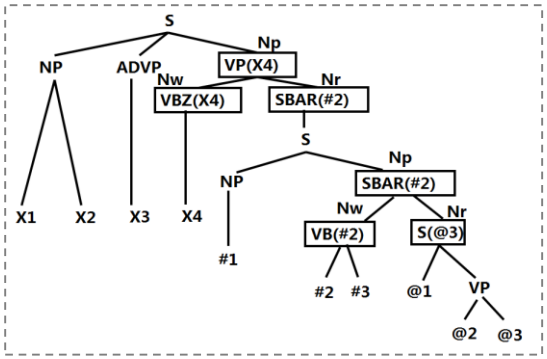
P(bi |,Wi, ,t) 表示Wi 是边界的概率 。 其中，bi ∈{ 边界，非边界}

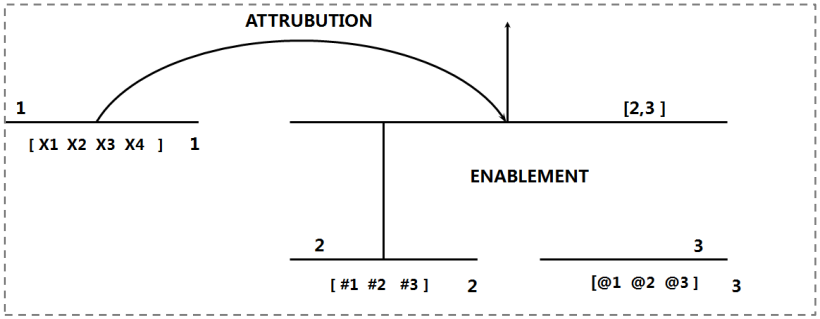
$$P(b|w,t) \simeq \frac{Cnt(N_p \rightarrow \ldots N_w \uparrow N_r \ldots)}{Cnt(N_p \rightarrow \ldots N_w N_r \ldots)}$$

训练语料 Penn Treebank (RST-DT, 2002) corpus；5809 triples of the form

$$\langle s, syntacticTree(s), discourseTree(s) \rangle$$

如：S: **X1 X2 X3 X4 #1 #2 #3 @1 @2 @3**





参数训练：最大似然估计 + （插值数据平滑）

13.1.2 基于RST的篇章分析法

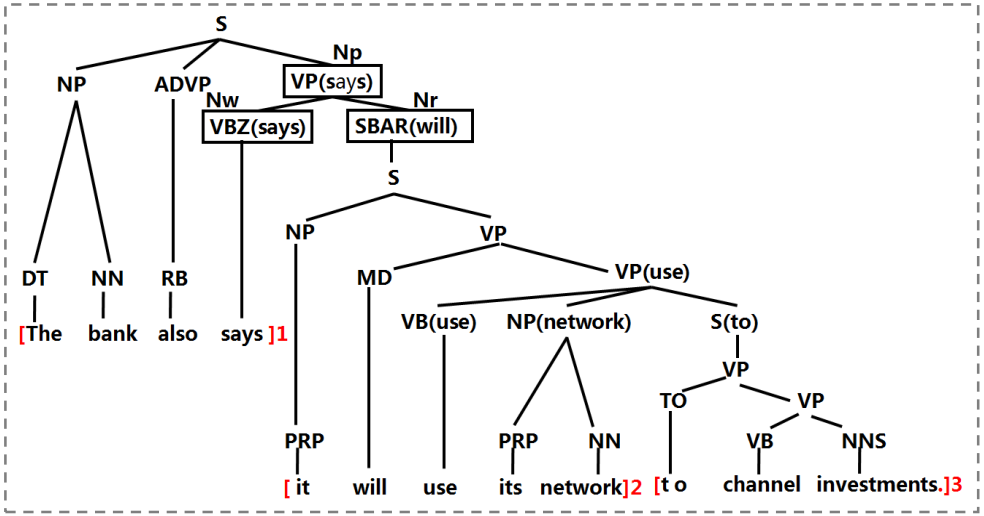
2.确定结构段目标

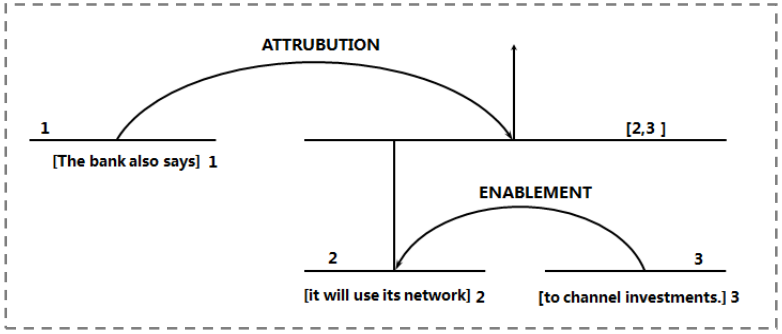
输入：

标注篇位边界的词汇化句法树（DS-LST）

输出：

篇章分析树





13.1.2 基于RST的篇章分析法

方法:

（1）. 定义篇章分析模型： 计算在给定参数下的 篇章分析树 DT 的概率

（2）. 篇章分析器：找篇章分析模型中概率最大的 分析树 DT_{best}

13.1.2 基于RST的篇章分析法

A. 篇章分析模型定义：

定义:

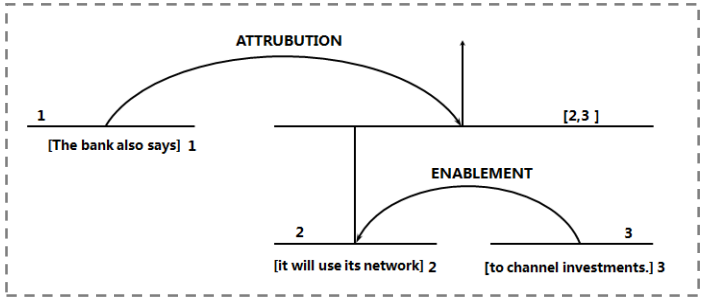
$$P(DT|\Theta) = \prod_{c \in DT} P_s(ds(c)|\Theta) \times P_r(rel(c)|\Theta)$$

模型结构

其中: DT为篇章分析树，θ 为 一组给定参数，c 为 篇章分析树的元组

ds(c) 为元组c篇章结构 ；rel(c)为元组c 篇章关系

名词解释:



篇章分析树元组 R (i,m,j)

R 篇位(i 到 m)与篇位（m+1到j）的 关系

R: 核心-卫星(NS), 卫星-核心S (SN), or 核心-核心 (NN).

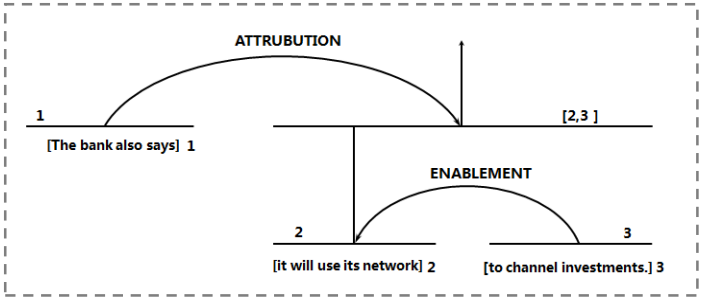
如: { ATTRIBUTION-SN[1,1,3]，ENABLEMENT-NS[2,2,3] }

13.1.2 基于RST的篇章分析法

P_s 表示篇章结构的 概率；P_r 表示篇章关系的概率

如： P_s

结构 (1, (2, 3)) > 结构((1,2), 3)



P_r

结构 (1, (2, 3))

ATTRIBUTION-NS > CONTRAST-NN

$$P(DT|\Theta) = \prod_{c \in DT} P_s(ds(c)|\Theta) \times P_r(rel(c)|\Theta)$$

问题: 如何确定 参数 θ ？

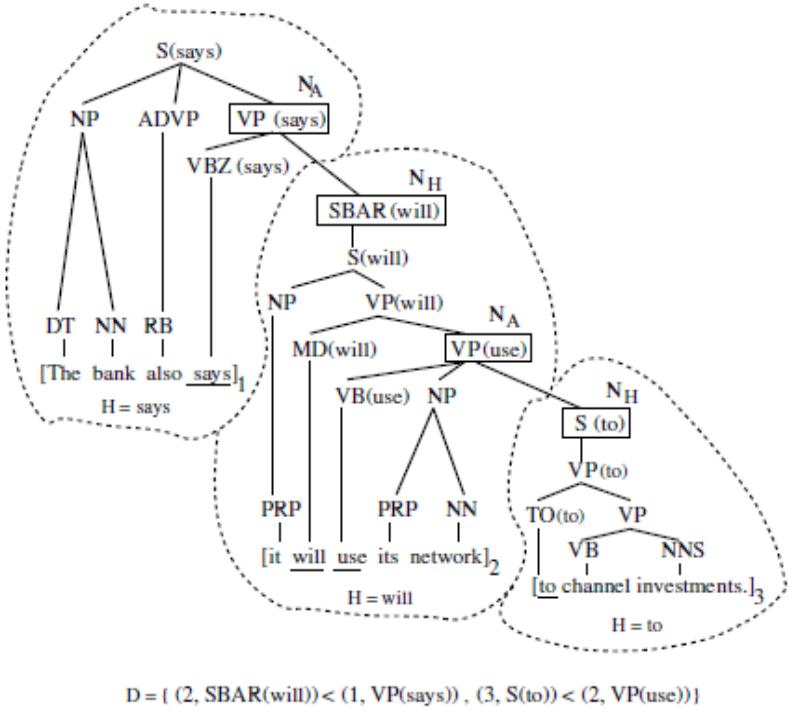
输入：DS-LST → DS-LST. 与 θ ？

13.1.2 基于RST的篇章分析法

定义DS-LST 支配集 （可提供θ 参数）

DS-LST 支配集：

定义每个EDUs 出现在DS-LST 中的头结点词 为 H，该词出现的最高结点为EDUs 头结点，定义为N_H.含有根结点的 EDU 例外，对于每个非例外EDU 的N_H均有父结点 N_A。且 N_H 与N_A 属于相邻的不同EDU。各EDU 由 N_H 与N_A 连接， 记为 $N_H < N_A$



DS-LST 支配集 为 其中所有$N_H < N_A$ 对组成。

如： Edu 2 H=will , N_H =SBAR(will) , N_A =VP(says)

Edu 3 H=to , N_H =S(to) , N_A =VP(us)

DS-LST 支配集D : { 2<1 , 3<2 }

13.1.2 基于RST的篇章分析法

支配集D 做 为

$$P(DT|\Theta) = \prod_{c \in DT} P_s(ds(c)|\Theta) \times P_r(rel(c)|\Theta)$$

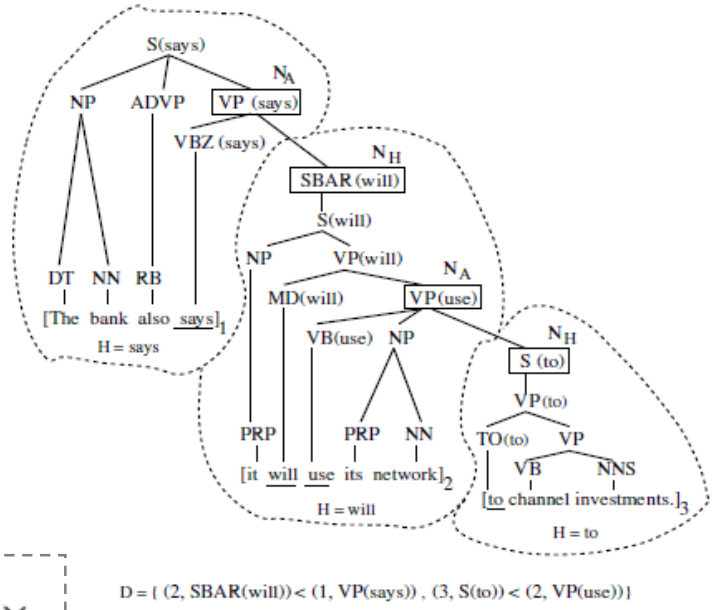
的条件参数θ,

方法 $$P(DT|\Theta) = \prod_{c \in DT} P_s(ds(c)|\Theta) \times P_r(rel(c)|\Theta)$$

转化为:

$$P(DT|D) = \prod_{c \in DT} P_s(ds(c)|filter_s(c, D)) \times P_r(rel(c)|filter_r(c, D)))$$

模型结构



其中,

Filter$_s$ 和 Filter$_r$ 作用是对每个元组只选择 在D中的信息作为结构概率和关系概率的计算条件;利用D可以 准确的确定 一个元组的 Ps 和 Pr

例如: 元组 c = ENABLEMENT-NS[2,2,3]

Filter $_s$ (c,D) ={(2,SBAR)<(1,VP),(3,S)<(2,VP)}

Filter $_r$ (c,D) ={S(to)<VP(use)}

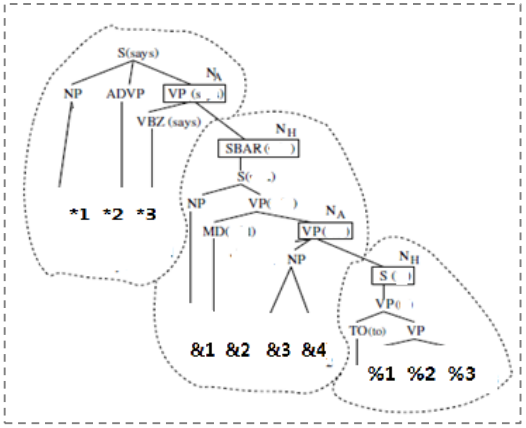
13.1.2 基于RST的篇章分析法

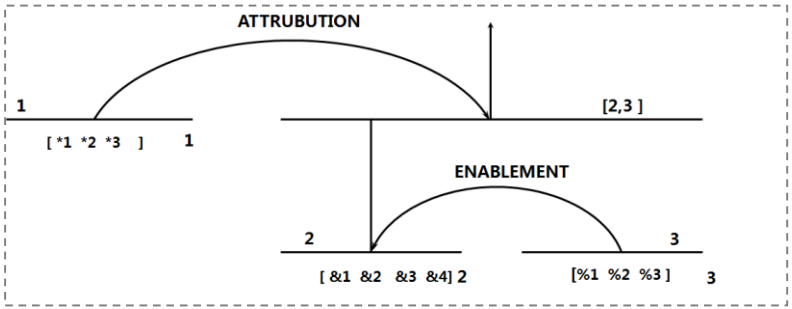
B.篇章分析模型参数学习

$$P(DT|D) = \prod_{c \in DT} P_s(ds(c)|filter_s(c,D)) \times P_r(rel(c)|filter_r(c,D)))$$

P_s 表示篇章结构的 概率；P_r 表示篇章关系的概率

训练语料



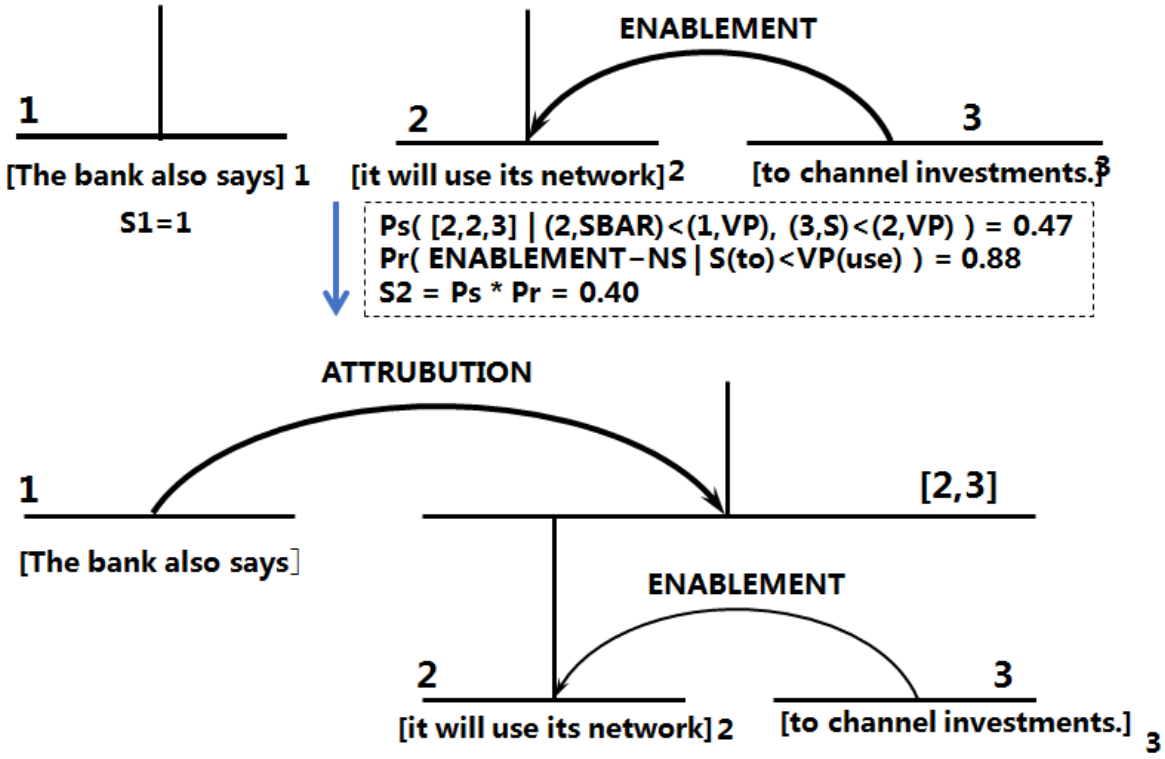


用最大似然估计

13.1.2 基于RST的篇章分析法

篇章分析器

自底向上 形成篇章分析树，计算每棵树的概率取概率最大的分析树为结果。



13.1.2 基于RST的篇章分析法

例2：

篇位1 ZERO POPULATION GROWTH
篇位2 November 22,1985
篇位3 Dear Friend of ZPG:
篇位4 At 7:00 a.m. on October 25, our phones started to ring.
篇位5 Calls jammed our switchboard all day.
篇位6 Staffers stayed late into the night, answering questions and talking with reporters from newspapers, radio stations, wire services and TV stations in every part of the country.
篇位7 When we released the results of ZPG’s 1985 Urban Stress Test, we had no idea we’d get such an overwhelming response.
篇位8 Media and public reaction has been nothing short of incredible!
篇位9 At first, the deluge of calls came mostly from reporters eager to tell the public about Urban Stress Test results and from outraged public officials who were furious that we had “blown the whistle” on conditions in their cities.
篇位10 Now we are hearing from concerned citizens in all parts of the country who want to know what they can do to hold local officials accountable for tackling population-related problems that threaten public health and well-being.
篇位11 ZPG‘s 1985 Urban Stress Test, created after months of persistent and exhaustive research, is the nation’s first survey of how population-linked pressures affect U.S. cities.
篇位12 It ranks 184 urban areas on 11 different criteria ranging from crowding and birth rates to air quality and toxic wastes.
篇位13 The Urban Stress Test translates complex, technical data into an easy-to-use action tool for concerned citizens, elected officials and opinion leaders.
篇位14 But to use it well, we urgently need your help.
篇位15 Our small staff is being swamped with requests for more information and our modest resources are being stretched to the limit.

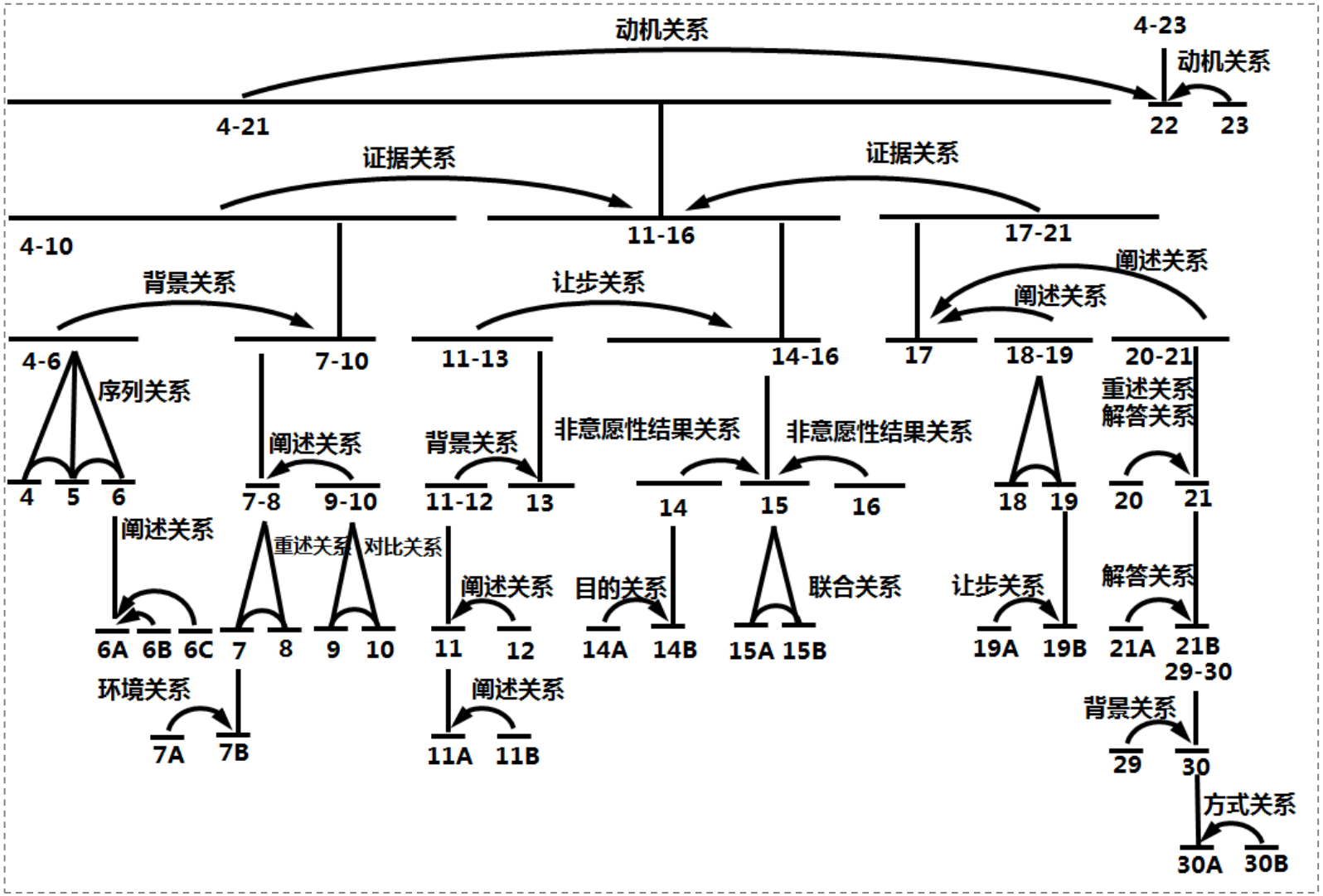
13.1.2 基于RST的篇章分析法

篇位16 Your support now is critical.
篇位17 ZPG’s 1985 Urban Stress Test may be our best opportunity ever to get the population message heard.
篇位18 With your contribution ZPG can arm our growing network of local activists with the materials they need to warn community leaders about emerging population-linked stresses, before they reach crisis stage.
篇位19 Even though our national government continues to ignore the consequences of uncontrolled population growth, we can act to take positive action at the local level.
篇位20 Every day decisions are being made by local officials in our communities that could drastically affect the quality of our lives.
篇位21 To make sound choices in planning for people, both elected officials and the American public need the population-stress data revealed by our study.
篇位22 Please make a special contribution to Zero Population Growth today.
篇位23 Whatever you give - $25, $50, $100 or as much as you can – will be used immediately to put the Urban Stress Test in the hands of those who need it most.
篇位24 Sincerely,
篇位25 (handwriting signature)
篇位26 Susan Weber
篇位27 Executive Director
篇位28 P.S.
篇位29 The results of ZPG’s 1985 Urban Stress Test were reported as a top news story by hundreds of newspapers and TV and radio stations from coast to coast.
篇位30 I hope you’ll help us monitor this remarkable media coverage by completing the enclosed reply form.

13.1.2 基于RST的篇章分析法

注：
6B:answering questions
6C:and talking with reporters ... Country
7B:we had no idea ... response.
11A:ZPG’s Urban Stress Test, ... Is the nations first survey ... cities
11B:created after months of persistent and exhaustive research.
14B:we urgently need your help.
15B:and our modest resources are being stretched to the limit.
19B:we can act to take positive action at the local level.
21:both elected officials ... study.
30B:by completing the enclosed

13.1.2 基于RST的篇章分析法



篇章分析树

13.1.2 基于RST的篇章分析法

RST理论的特点

1.语篇核心

RST分析的一个重要成果就是确定了体现语篇要旨的核心篇位,这一篇位称为综合效果位置。

2.关联词语与关系命题

关系结构本身也可以象小句结构和词语一样表达意义和信息。这种表达可以有各种形式标记(关联词语或形态、句法特征等),也可以完全是隐含的。 RST认为语篇的连贯性来源于关系结构而不是表层的词汇或语法标记。

13.1.2 基于RST的篇章分析法

3 从属连接

传统的“从属连接”应看作是语篇中普通存在的“核心一辅助结构段”关系模式在分句层的语法化, 并应取消“从属连接”的提法,而具体区分“**主从连接**”(hyPotaxis)与“**内嵌**”(embedidng)。主从连接只包括传统的“从属连接”中的时间、原因、条件.结果等状语从句: 而内嵌则主要包括限制性定语从句,主语和宾语从句及动词和形容词的补语从句。

13.1.2 基于RST的篇章分析法

RST理论的优势

采用功能主义的方法,摆脱了关联词语、语法结构、类型结构等表层语言形式的束缚,通过对语篇的深层结构一关系结构的阐述揭示语篇功能,具有较强的解释力。比较令人信服地说明了语篇的整体性及连贯性来源于内部结构关系,而不是表层的某种连接模式

语篇中大量存在的无标记隐含关系命题是形式主义的语篇分析理论一直无法解释的难题,RST理论从功能的角度为这一现象提供了解释框架,这对于从根本上揭示语篇的连贯性及交际功能具有重要意义。

RST理论的不足

RST理论的描写范围仍局限于非对话体的书面语篇,对口语体,对话体语篇无法适用。

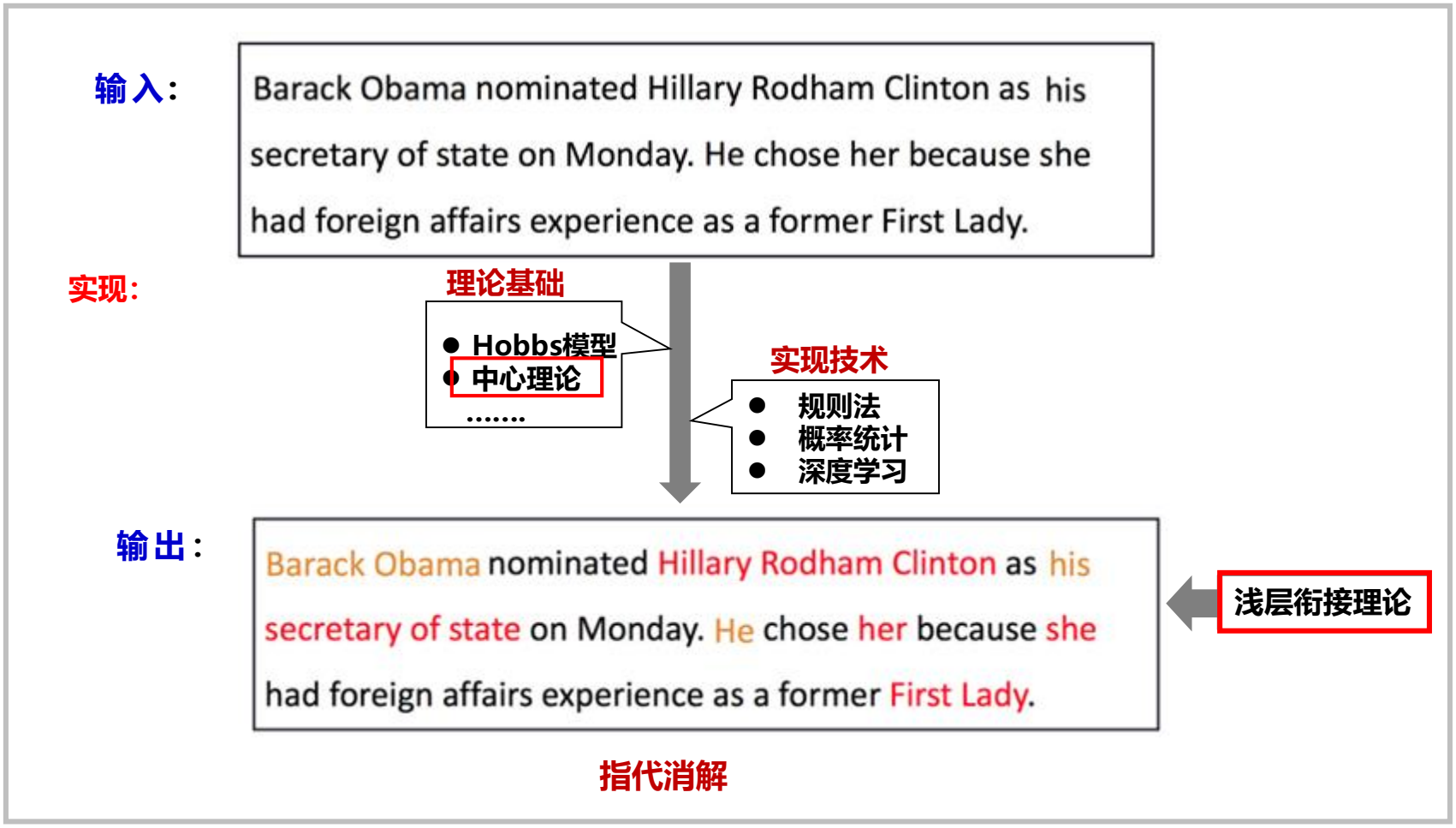
内 容 提 要

13.1　篇章连贯性分析

13.2　篇章衔接性分析

篇章衔接性概述

基于浅层衔接理论的篇章衔接性分析任务



内 容 提 要

13.1 篇章连贯性分析

13.2 篇章衔接性分析

13.2.1 浅层衔接理论

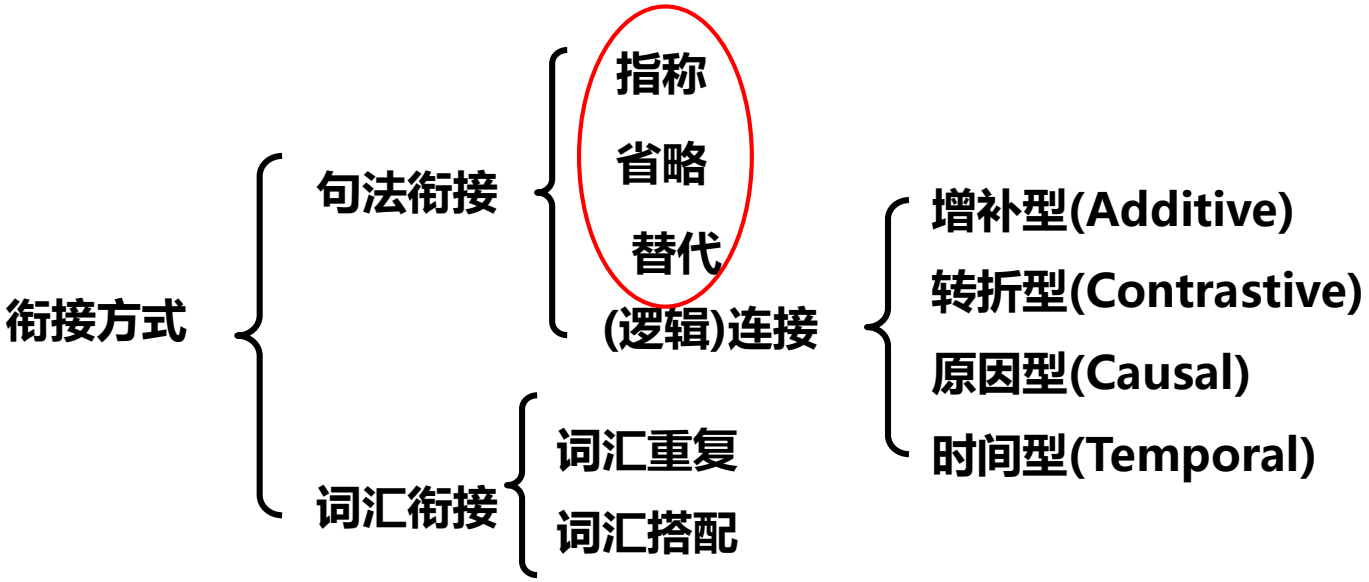
13.2.2 中心理论

13.2.3 指代消解方法

13.2.1 浅层衔接理论

浅层衔接理论

Halliday 的浅层**衔接理论**是最早研究篇章衔接关系的理论体系 。浅层衔接理论指出“**当篇章中的某个成分的解释依赖于篇章中另一个成分的解释时，这两个成分之间就产生了衔接关系**”。



13.2.1 浅层衔接理论

句法衔接

1.人称代词((Pronoun)

例1.[李明i]怕[高妈妈j]一人呆在家里寂寞，[他i]便将家里的电视搬了过来。

2.指示代词(Demonstrative)

例2.[很多人都想创造一个美好的世界留给孩子i] ,[这i]可以理解，但不完全正确

3.有定描述(Definite Description)

例3. [贸易制裁i]仿佛成了[美国政府在对华关系中惯用的大棒i]，然而，[这根大棒i]果真如美国政府所希望的那样灵验吗?

4. 也有些指示语没有标记

例4[沈阳矿山机器集团公司i] 的领导在创造力开发活动中大胆创新，…，有效遏制住了经济滑坡，「公司i] 产值以平均每年33%的幅度递增。

NLP中为了能够理解篇章的整体含义需要对这种衔接关系进行“指代消解”

13.2.1 浅层衔接理论

指代消解

指代：篇章中的一个语言单位（通常是词或短语）与之前出现的语言单位存在特殊语义关联，其语义解释依赖于前者。

如：　**李明** 怕高妈妈一人呆在家里寂寞，**他** 便将家里的电视搬了过来。

人们都想创造美好的世界留给孩子，**这**可以理解，但不完全正确

在语言学把用于指向的语言单位(抽象的语言单元)称为照应语（或指代语 Anaphor），被指向的语言单位（具体的实体）称为先行语（或先行词 Antecedent）

指代消解：确定照应语所指的先行语的过程就是指代消解

13.2.1 浅层衔接理论

指代消解的分类

根据语言学知识**从照应语的角度**把指代消解分为三类

(1) 按先行词与照应语出现的顺序分类

在篇章中当照应语的位置在先行语之前则称为**预指消解**.当照应语位于先行语之后称为**回指消解**.

(2) 按照应语的抽象程度分类

根据指代的表现形式的抽象程度，指代消解分为**名词消解**、**代词消解**、**零代词消解**. 零代词是指篇章中句子成分缺省的部分是在前文中提起的某个实体.零代词在中文句子中出现的频率很高。具体有6类：

- Indefinite NPs（无定名词）：一辆汽车
- Definite NPs （有定名词）：那个人
- Pronouns （人称代词）：它，他
- Demonstratives （指示代词）：这，那
- One-anaphora （one指代）：one (in English)
- Zero anaphora （0型指代）：省略

13.2.1 浅层衔接理论

Indefinite NPs （无定名词）

- 为读者引入一个新的实体时常用无定形式；
- 引入的实体，可能的确存在（明确的），也可能不明确；
- 如： – 张先生娶了一位法国太太(Specific)
  – 史密斯想娶一位中国姑娘(non-specific)

Definite NPs （有定名词）：

- 无论读者知道否，一定存在。
- 如： –首位进入太空的宇航员（即，前苏联宇航员尤里.加加林）；
  (通过某些知识可以知道)
  –Look, how beautiful the girlis! (实际存在)
  特点：定冠词（这/那）引导的名词短语

13.2.1 浅层衔接理论

Pronouns （人称代词）：

- 典型的指示代词包括：他 、它,…
- 用于指示前面提及的人。
- 如: 刘博士刚买了一辆车，他去加油。

Demonstratives （指示代词）

- 典型的指示代词包括：那, 这,…
- 当指示代词与后面的名词(短语)连用时，此时变为了定冠词，形成有定表示.
- 如: 刘博士刚买了一套房子，那是一套性价比相当好的房子。

13.2.1 浅层衔接理论

One-anaphora （one指代）

- 出现在英语中
- 表示某集合中的一个元素.

如：–He had a BMW before, now he got another one.
–John has two BMWs, but I have only one.

Zero anaphora （0型指代）

- **0形式的判断：**

需要在句子层面上判断哪些必须的成分省略了

如：(1)美国宣布（X）部分取消（X）对朝鲜长达近半个世纪的经济制裁。

(2) 李向阳机智地组织游击队攻城并烧毁了敌人的粮库，（X）迫使松井撤出了李庄。

- **0形式恢复（消解）**

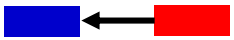
13.2.1 浅层衔接理论

(3) 按照应语在句子中语义关系强弱程度分类

当先行语和照应语存在等价关系，并同时指向同一个实体时叫做**共指**. 共指关系在脱离上下文的语义仍旧独立存在，与上下文关系较弱.

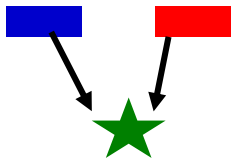
(非等价) **指代消解**是指先行语与照应语之间存在着非对称的关系并且和上下文的语义有着紧密联系，在不同的语义和语境下照应语指代的先行语是不同的.

如： **李明** 怕高妈妈一人呆在家里寂寞，**他**便将家里的电视搬了过来。



回指

香港首任行政长官董建华出席了会议。



共指

两者的目标：

- 指代消解：寻找指代语对应的先行语
- 共指消解：发现指向相同实体的语言表示单元（包括多语篇）

13.2.1 浅层衔接理论

指代消解分类小节：

与先行语在句子中位置顺序		表达抽象程度		句子中语义关系强弱程度	
预指	前（照应语在前）	零代词	弱	等价（共指）	弱
	↓	代词	↓		↓
回指	后（照应语在后）	名词短语	强	非等价	强

指代消解依照应语进行分类

内 容 提 要

13.1　篇章连贯性分析

13.2　篇章衔接性分析

13.2.1 浅层衔接理论

13.2.2 中心理论

13.2.3 指代消解方法

13.2.2 中心理论

中心理论（Centering Theory）

Grosz and Sidner (1983)创立,是一种关于语篇结构的理论。

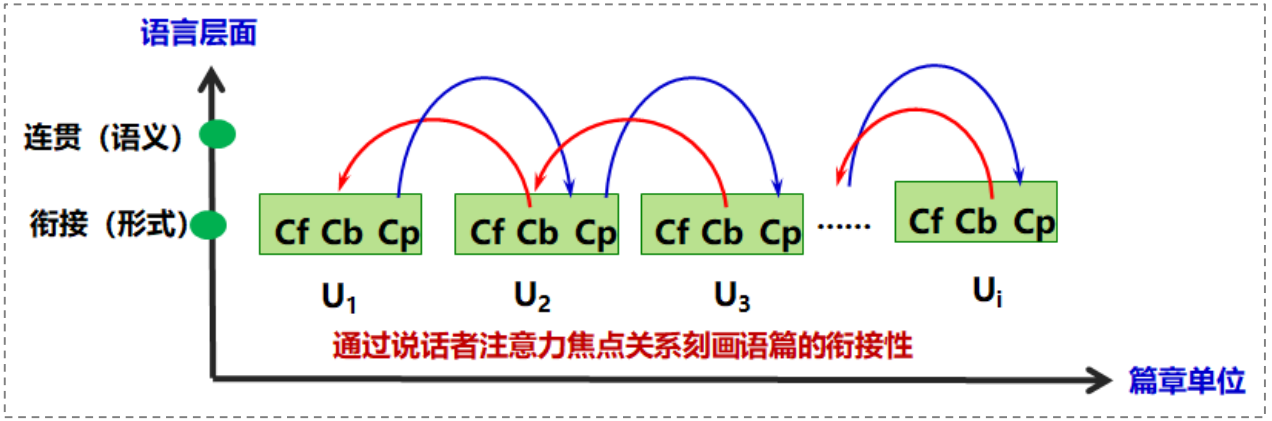
该理论认为篇章由三个分离的但相互联系的部分组成：**话语序列结构**（语言结构），**目的结构**（说话者意图）和 **关注焦点状态**（说话者注意力状态）

中心理论对关注状态进行模型化，将关注焦点描述为“中心”

通过说话者注意力焦点来阐述 语篇的衔接性

13.2.2 中心理论

中心理论



要素：

- **中心**：Cf:前看中心， Cb:回视中心， Cp:优先中心
- **话题关系**：根据**回视中心的变化状态**用毗连着的语句关系来界定语篇结构的衔接性

特点： 描述的话题结构呈线性,不区分复杂的等级层次;最大的优势是把复杂的话题关系量化,把话题间关系的判断公式化,可操作性强 。

13.2.2 中心理论

中心 (center)：话语中的语义实体,它通常是名词性的。

中心理论： 每个语篇单位有三个中心

中心
- 前看中心 (Cf.)　序列列表（有序）
- 优先中心 (Cp.)　唯一
- 回视中心(Cb ）　唯一

- **前看中心 (forward-looking center list,Cf.)**

一个话语单元（utterance) 通常包含若干个中心,它们根据语法关系的显著性和从左到右出现的线性顺序,形成一个中心序列，称为前看中心(Cf.)

如：（1） Cooper is standing around the corner

（2） He is waiting for Grey

对于（1） 前看中心： Cooper， corner

对于（2） 前看中心： He， Grey

13.2.2 中心理论

- **优先中心(preferred center, Cp.)**

  前看中心序列中排列第一的成分

如：（1） Cooper is standing around the corner

（2） He is waiting for Grey

对于（1） 前看中心： Cooper， corner

优先中心： Cooper

对于（2） 前看中心： He， Grey

优先中心： He

13.2.2 中心理论

- **回视中心(backward-looking center,Cb.)**

同时出现在当前和前一个分析单元中,且排序相对最靠前的那个中心

如: （1） Cooper is standing around the corner

（2） He is waiting for Grey

对于（1） 前看中心： Cooper， corner

优先中心： Cooper

回视中心： NULL

对于（2） 前看中心： He， Grey

优先中心： He

回视中心： He=Cooper

13.2.2 中心理论

话题关系

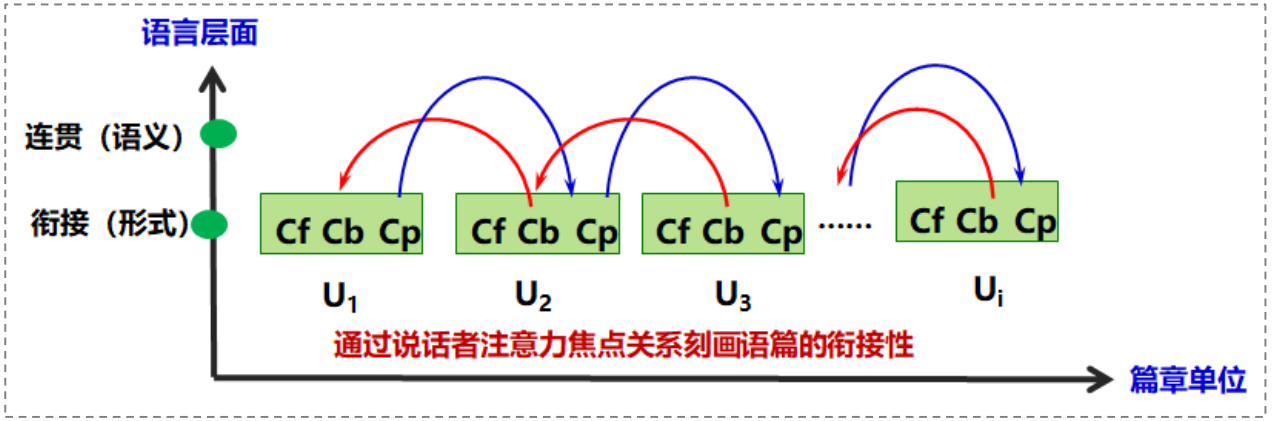
中心理论根据 **回视中心的变化状态** 将毗连着的语句关系分为四种,并由此来界定语篇结构的衔接性

中心理论话题关系主要有四种:

- 延续话题(continue)
- 保持话题(retain)
- 顺畅度转换(smooth shift)
- 不顺畅转换(rough shift)

13.2.2 中心理论

话题关系类型定义



	$Cb(U_i)=Cb(U_{i-1})$	$Cb(U_i)\neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	延续	顺畅转换
$Cb(U_i)\neq Cp(U_i)$	保持	不顺畅转换

$Cb(U_i)$=当前语句回指中心

$Cp(U_i)$=当前语句优选中心

$Cb(U_{i-1})$=上一语句回指中心

话题关系优先级：

延续话题>保持话题>顺畅转换>不顺畅转换

13.2.2 中心理论

例： 用语篇单元间的衔接关系对下面用两个语段对上述观点略加分析:

a. David loved Elizabeth. He had known her for years. At one time he had disliked her. She, on the other hand, hated him . She had always thought he was a creep.

b. David loved Elizabeth. She, on the other hand, hated him. He had known her for years. She had always though the was a creep . Atone time he had disliked her.

13.2.2 中心理论

	$Cb(U_i)=Cb(U_{i-1})$	$Cb(U_i)\neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	延续	顺畅转换
$Cb(U_i)\neq Cp(U_i)$	保持	不顺畅转换

分析： **a**.

1. David loved Elizabeth.

   Cf={David, Elizabeth}；Cb= NULL

2. He had known her for years.

   Cf={He（David）, her(Elizabeth)}; Cb= David (延续)

3. At one time he had disliked her.

   Cf={He（David）, her(Elizabeth)}; Cb= David (延续)

4. She, on the other hand, hated him .

   Cf={She(Elizabeth),him(David)}; Cb = Elizabeth(顺畅转换)

5. She had always thought he was a creep.

   Cf={She(Elizabeth),him(David)}；Cb= Elizabeth(延续)

a. 的语篇结构进展连贯,过渡流畅:(回指)中心延续(句1—句3)+中心转换(句4)+中心延续(句5)。

13.2.2 中心理论

	$Cb(U_i)=Cb(U_{i-1})$	$Cb(U_i)\neq Cb(U_{i-1})$
$Cb(U_i)=Cp(U_i)$	延续	顺畅转换
$Cb(U_i)\neq Cp(U_i)$	保持	不顺畅转换

分析： **b.**

1. David loved Elizabeth.

   Cf={David, Elizabeth}；Cb= NULL

2. She, on the other hand, hated him.

   Cf={She(Elizabeth) , him(David) }；Cb= Elizabeth

3. He had known her for years.

   Cf={He(David), her(Elizabeth)}；Cb= David (转换)

4. She had always though he was a creep .

   Cf={She(Elizabeth) , he(David) }；Cb= Elizabeth(转换)

5. At one time he had disliked her.

   Cf={He(David), her(Elizabeth)}；Cb= David (转换)

b. 的语句之间表征的只有中心转换状态,毫无连贯性可言。

显然语篇 a在结构上比语篇 b 流畅（衔接性好）

13.2.2 中心理论

中心之间替换:

如果在相邻的两个分析单元中,出现了**语义上相关**,但是又**有区别的中心**,把这些中心进行**恰当地替换,**可以让他们之间的关系更明朗,从而使话题之间的关系判断更明晰。比如:一般代词被所指称的实际名词替换;同义词之间的替换;上义词与其下义词之间的替换;整体与其部分间的替换(Hadic&Taboada,2006)

运用这种替换技巧可以进行指代消解

13.2.2 中心理论

中心理论的局限性：

- 对篇章中心的刻画只能考虑局部的连贯性，没有对全局的连贯性加以考虑，所以消解工作只限于相邻的句子;
- 主要用于人称代词消解，对零指代以及名词短语的消解效果不好;
- 当需要指代的部分较多时很难做出准确判断.

内 容 提 要

13.1　篇章连贯性分析

13.2　篇章衔接性分析

13.2.1 浅层衔接理论

13.2.2 中心理论

13.2.3 指代消解方法

13.2.3 指代消解方法

指代消解方法:

- **基于规则的方法**
  - **基于句法结构的方法**：Hobbs 提出的 hobbs 算法
  - **基于语篇结构的方法**：Brennan等提出了一种基于中心理论的代词消解算法即 BFP算法
- **概率统计方法**
  - **基于有监督的指代消解方法**
  - **基于无监督的指代消解**
- **深度学习方法**

13.2.3 指代消解方法

★基于中心理论的代词回指解析计算法

基于中心理论 的代词指代消解规则

- 如果Cf(ui-1) 的某元素以代词形式出现在ui, 那么，这个元素就是Cb(ui)

  规则给出了凸显性的直观解释，即被代词表示的实体具有显著性
- 如果有多个代词，那么其中之一是Cb(ui)
- 如果只有一个代词，那么一定是Cb(ui)

Cb(ui)的确定依赖于两个条件:

(1) 一定是在Ui中出现的语义实体;

(2) 该实体也一定在Cf(Ui-1)中出现过，如果Ui有多个实体也在Ui-1中出现，那么，作为Cb(Ui)的实体在Cf(Ui-1)中应有更高的排位。

根据代词，篇章的读者或者听者可以保持注意力。对于篇章的作者，不使用代词会使得篇章的流利度降低。

13.2.3 指代消解方法

算法BFP(Brennan, Friedman and Pollard,1987-**代词消解**）

算法思想:

Step1. 如果在U i 中出现**人称代词**，则自左至右顺序检验Cf(U i -1)中的元素，直至同时满足词汇句法（Morphosyntactic）、约束（Binding）和类型标准（Sortal criteria）；这样的元素作为先行语；

Step2. 完全读取表述Ui ， 生成Cf(Ui)，对Cf(Ui)进行排序, 计算Cb(Ui)

13.2.3 指代消解方法

例1. a) The sentry was not dead.

Cb:-

Cf:[Sentry]

(b) He was in fact, showing signs of reviving...

Cb: he (he=Sentry)

Cf:[he（Sentry）, signs]

(c) He was partially uniformed in a cavalry tunic.

Cb: he (he=Sentry)

Cf:[he （Sentry）, tunic]

(d) Mike stripped this from him and donned it.

Cb: him (him=Sentry)

Cf:[Mike, this, it, him （Sentry）]

(e) He tied and gagged the man.

Cb: he=Mike,

13.2.3 指代消解方法

例2： 如下三个句子构成简单篇章：

（1） Cooper is standing around the corner

（2） He is waiting for Grey

（3） He intends to see film with him

使用中心理论对这段话进行分析可以得到如下结果：

（1） Cooper is standing around the corner

回视中心： NULL

前看中心： Cooper ， corner

优先中心： Cooper

（2） He is waiting for Grey

回视中心： He = Cooper

前看中心： He（Cooper） ， Grey

优先中心： He

状态转换： 连续

13.2.3 指代消解方法

(3) He intends to see film with him

情况一

回视中心： He = Cooper , him = Grey

前看中心： He， film ,him

优先中心： He

状态转换： 连续

情况二

回视中心： He = Grey, him =Cooper

前看中心： He， film ,him

优先中心： He

状态转换： 转换

根据优先规则 连续>转换　　情况一为分析结果

参考文献：

王伟，“修辞结构理论”评介

Soricut and Marcu，Sentence Level Discourse Parsing using Syntactic and Lexical Information

宗成庆，统计自然语言处理（第2版）课件

夏志华，中心理论—话题与韵律接面研究的新方法

周炫余，刘娟等，篇章中指代消解研究综述

王厚峰，语篇分析与指代消解 ，课件

在 此 表 示 感 谢!

信息内容安全技术国家工程实验室
National Engineering Laboratory for Information Security Technologies

谢谢各位！