

篇章（内）的计算

目录

- ▶ **语篇分析简介**
- ▶ 指代和指代消解
- ▶ 衔接和连贯
- ▶ 篇章表示和相似度计算

自然语言处理的不同层次

- Morphology
 - 词的构成问题
- Syntax(Parsing)
 - 词与词之间的结构关系
- Semantics
 - 词的意义、词与词组合（短语/句子）意义
- Discourse
 - 句子之间的关系，上下文的意义。

语篇 (Discourse)

- 前后意义关联的句子序列。
- 几种说法：
 - 话语、语篇、篇章、文本 (英文: **discourse, text**)
- 两个例子：
 - Ex1: 比尔来自于美国。今天交通非常拥挤。长江贯穿中国的多个省市。因此，计算语言学是计算机科学与语言学的交叉。
4 correct sentences but collectively do not make meaning
 - Ex2: 这里的交通非常拥挤。张先生早上**6: 40**之前就得出发去上班，常常会提前半小时到单位；如果稍晚一点，他就很可能迟到。
it makes meaning

目录

- ▶ 语篇分析简介
- ▶ **指代和指代消解**
- ▶ 衔接和连贯
- ▶ 篇章表示和相似度计算

关于指代

- 指代现象的产生

- 假设有这样一组句子：

张三一大早就赶到了学校。张三先到食堂吃早餐，然后张三到张三的宿舍拿张三自己的教材和张三自己的笔记本。当张三匆忙来到教室时，张三发现张三的课本拿错了。

- 设想修改为这样表达：

张三一大早就赶到了学校。他先到食堂吃早餐，然后[X]到[X]宿舍拿自己的教材和[X]笔记本。当[X]匆忙来到教室时，他发现[X]课本拿错了。

- 哪一种表达更符合人们的习惯？

- 语言的表达追求“经济”与“变化”

指代的定义

- 指代(anaphora) 的定义(Hirst, 1981) :

指代语 (Anaphor) 先行语 (Antecedent)

Anaphora is the device of making in discourse an abbreviated **reference** to some **entity** in the expectation that the perceiver will be able to disabbreviate the reference and thereby determine the identity of the entity.

六类指称表示

- ▶ Indefinite NPs（不定名词）：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

Indefinite NPs

- ▶ 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;
- ▶ 两个例子:
 - 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

- 无论读者知道否，一定存在
 - 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 girl** is! (实际存在)
 - 为了消除小士兵对生人的陌生感，两位女记者带着**这个小男孩**逛街...(在上下文中)
- 已经在上下文中出现时，需要指代消解
 - 特点：定冠词（这/那）引导的名词短语

Demonstratives

- 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
- **Ex:** 刘博士刚买了一套房子，**那**是一套性价比相当好的房子。

One-anaphora (替换)

- 出现在英语中
- 表示某集合中的一个元素.
- Ex:
 - He had a BMW before, now he got another **one**.
 - John has two BMWs, but I have only **one**.

省略 - 零指代 (Zero anaphora)

- 一个例子
 - 张三一大早就赶到了学校。**他**先到食堂吃早餐，然后[X]到宿舍拿自己的教材和[X]笔记本。当[X]匆忙来到教室时，**他**发现[X]课本拿错了。
- 英语中的零指代很少见，但汉语中十分常见：
 - They said **they** were coming to help us with **our** house repair today.
 - 他们说[X]今天来帮我们修[X]房子
 - 他们说**他们**今天来帮我们修**我们的**房子（很少这样说）

回指与共指

- ▶ 指代一般包括两种情况：
 - 前指(Anaphora): 强调指代语与另一个表述之间的关系。指代语的指称对象通常不明确，需要确定其与先行语之间的关系来解释指代语的语义
 - 张先生走过来，给大家看他_他的新作品
 - 共指(coreference): 强调一个表述与另一个表述是否指向相同的实体，可以独立于上下文存在
 - 第44任美国总统 与 奥巴马
- ▶ 有时候回指和共指并没有严格的区分，可以简单的统称为指代

指代消解问题

- ▶ 回指消解：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
- ▶ 可用的解决方案：
 - ▶ 中心理论
 - ▶ 分类方法
 - ▶ 等等

中心理论 (Centering Theory)

- ▶ 语篇由不同的语段组成
 - ▶ 通常用 U_i 表示语段， $\{U_1, U_2, \dots, U_n\}$ 表示 n 个语段组成的篇章
- ▶ 中心（center）：在篇章片段中，联系不同语段的实体
 - ▶ 一般用 C 表示中心

中心的类型

- ▶ **前看中心 (forward-looking center):** 当前话语中所提及的名词性实体
 - ▶ 表示一个语段可能存在的会话焦点，可能会有多个
 - ▶ 通常用 $C_f(U_i)$ 表示，是一个有序列表
 - ▶ 排序关系：主语 > 直接宾语 > 间接宾语 > 其他实体
- ▶ **回看中心 (backward-looking center):** 前看中心的特殊成分，表示当前话语所谈论的中心
 - ▶ 只能有一个实体，通常用 $C_b(U_i)$ 表示
- ▶ **优选中心 (Preferred Center):** $C_f(U_i)$ 中排序最靠前的实体
 - ▶ 只能有一个实体，通常用 $C_p(U_i)$ 表示

回看中心的确定

- 回看中心的构造规则：
 - 如果 $C_f(U_{i-1})$ 的某实体以代词形式出现在 U_i , 那么, 这个元素就是 $C_b(U_i)$
 - 如果有多个代词, 那么其中排序最为靠前的是 $C_b(U_i)$
 - 如果只有一个代词, 那么一定是 $C_b(U_i)$
- 解释: $C_b(U_i)$ 的确定依赖于两个条件:
 - (1) 一定是在 U_i 中出现的语义实体;
 - (2) 该实体也一定在 $C_f(U_{i-1})$ 中出现过, 如果 U_i 有多个实体也在 U_{i-1} 中出现, 那么, 作为 $C_b(U_i)$ 的实体在 $C_f(U_{i-1})$ 中应有更高的排位。

中心理论：示例

语段		中心
U_1	小明和妈妈去逛街。	C_f : 小明、妈妈、街 C_p : 小明 C_b : Null
U_2	他看中了一件衣服。	C_f : 小明（他）、衣服 C_p : 小明（他） C_b : 小明（他）
U_3	但小明太胖了，	C_f : 小明 C_p : 小明 C_b : 小明
U_4	所以妈妈没给小明买那件衣服。	C_f : 妈妈、衣服、小明 C_p : 妈妈 C_b : 小明

基于中心理论的指代消解算法

1. 为每个语段中的实体生成可能的 C_b 、 C_f 组；
2. 通过各种约束条件来过滤（比如：句法位置约束、语义选择限制，等等）；
3. 通过连贯性来评级：如果一个代词R的指代成分为A所得到的篇章连贯性高于指代成分为B时得到的篇章连贯性，则将R的指代成分确定为A。

如何比较连贯性的高低？

中心转换关系

$$C_b(U_i) = C_b(U_{i-1}) \quad C_b(U_i) \neq C_b(U_{i-1})$$

或 $C_b(U_{i-1}) = \text{Null}$

$$C_b(U_i) = C_p(U_i)$$

CONTINUING

SMOOTH SHIFT

$$C_b(U_i) \neq C_p(U_i)$$

RETAINING

ROUGH SHIFT

- 连贯性比较 ***CON > RET > SSH > RSH***

连贯性比较示例

U_1 . John went to his favorite music store to buy a piano.

U_2 . He had frequented the store for many years.

U_3 . He was excited that he could finally buy a piano.

U_4 . He arrived just as the store was closing for the day.

U_1 . John went to his favorite music store to buy a piano.

U_2 . It was a store John had frequented for many years.

U_3 . He was excited that he could finally buy a piano.

U_4 . It was closing just as John arrived.

- 由中心理论推断：“第一段比第二段连贯”

第一段的中心转换

- U_1 . John went to his favorite music store to buy a piano.
 $C_f(U_1) = (\text{John}, \text{store}, \text{piano})$. $C_p(U_1) = \text{John}$
- U_2 . He had frequented the store for many years.
 $C_b(U_2) = \text{John}$. $C_f(U_2) = (\text{John}, \text{store})$. $C_p(U_2) = \text{John}$

CONTINUATION

- U_3 . He was excited that he could finally buy a piano.
 $C_b(U_3) = \text{John}$. $C_f(U_3) = (\text{John}, \text{piano})$. $C_p(U_3) = \text{John}$

CONTINUATION

- U_4 . He arrived just as the store was closing for the day.
 $C_b(U_4) = \text{John}$. $C_f(U_4) = (\text{John}, \text{store})$. $C_p(U_4) = \text{John}$

CONTINUATION

第二段的中心转换

- U_1 . John went to his favorite music store to buy a piano.
 $C_f(U_1) = (\text{John, store, piano})$. $C_p(U_1) = \text{John}$
- U_2 . It was a store John had frequented for many years.
 $C_b(U_2) = \text{John}$. $C_f(U_2) = (\text{store, John})$. $C_p(U_2) = \text{store}$
RETAINING.
- U_3 . He was excited that he could finally buy a piano.
 $C_b(U_3) = \text{John}$. $C_f(U_3) = (\text{John, piano})$. $C_p(U_3) = \text{John}$
CONTINUATION.
- U_4 . It was closing just as John arrived.
 $C_b(U_4) = \text{John}$. $C_f(U_4) = (\text{store, John})$. $C_p(U_4) = \text{store}$
RETAINING.

算法例子

U_1 : John saw a beautiful Acura Integra at the dealership.

U_2 : He showed it to Bob.

U_3 : He bought it.

指代消解问题:

U_2 : he = ? it = ?

U_3 : he = ? it = ?

例子分析 (1)

语段		中心	跳转类型
U_1	John saw a beautiful Acura Integra at the dealership.	C_f : John、Integra、dealership C_p : John C_b : Null	无
U_2	He showed it to Bob.	C_f : He=John、it={Integra, dealership}、Bob C_p : John C_b : John	continue
U_3	He bought it.	C_f : He=John 、it={Integra, dealership} C_p : John C_b : John	continue

例子分析 (2)

语段		中心	跳转类型
U_1	John saw a beautiful Acura Integra at the dealership.	C_f : John、Integra、dealership C_p : John C_b : Null	无
U_2	He showed it to Bob.	C_f : He=John、it={Integra, dealership}、Bob C_p : John C_b : John	continue
U_3	He bought it.	C_f : He=Bob 、it={Integra, dealership} C_p : Bob C_b : Bob	smooth

基于分类（ML）的指代消解

- 利用机器学习方法建立分类器：
- 方法：
 - 选取对共指消解产生影响的特征，主要包括：
 - 两者的距离, 字符的匹配程度, 单复数一致性, 性别一致性, 语义类的一致性, 是否是别称....
 - 例子：

[**聂/nr** **卫平/nr**] 今天/t 取胜/v 不易/a 。/w 布局/vn 阶段/n 便/d 与/p
[**实力派/n** **人物/n**] [**刘/nr** **小光/nr**] 九/m 段/q 展开/v 激战/vn , /w 棋
局/n 跌宕起伏/l , /w 互/d 有/v 优劣/n 。/w 直到/v 官子/vn 阶段
/n , /w [**聂/nr** **卫平/nr**] 才/d 因/c [**对手/n**] 的/u 缓/a 手/n 而/c
最终/d 取胜/v 。/w

[**对手/n**] => [**聂/nr** **卫平/nr**] 属于一类吗 ?

[**对手/n**] => [**刘/nr** **小光/nr**] 属于一类吗 ?

指代消解的应用

- ▶ 基本上文本处理相关的一切应用都需要用到指代消解
- ▶ 机器翻译：
 - ▶ **They** 是翻译成“他们”，“她们”，还是“它们”？
- ▶ 文本摘要：
 - ▶ 避免名字（同一个词）的反复使用，用代词（或0-形式）表示，以便符合习惯
- ▶ 实体链指：
 - ▶ 实体常常用代词表示，关系的建立需要明确代词的指向

目录

- ▶ 语篇分析简介
- ▶ 指代和指代消解
- ▶ **衔接和连贯**
- ▶ 篇章表示和相似度计算

意义相关性

- ▶ **语篇**：前后意义相关的句子序列
- ▶ 语篇应该是“合理”的，所有的句子应当围绕某个话题或中心展开，具有语义上的相关性
- ▶ 语篇应该是“简洁易懂”的，不仅在语义上相关，还应该在形式上关联

意义相关性的体现（1）

- 例子：
 - 张三擅长素描。他给家里的每个人都画了一幅[]，挂在房间的[]是自画像。
- 意义上是如何关联的？
 - 通过词汇语义表达关联：
 - 围绕着“画”而展开：素描、画像、一幅[]
 - 通过“指代”形成关联
 - ’ 人称代词“他”；
 - ’ 零型代词[]所表示的对象
 - 以词汇表示的关联，通常称为“衔接(cohesion)”

意义相关性的体现（2）

- 例子：
 - [s1]张三把李四的车钥匙藏起来了。[s2]他喝醉了。
 - [s3]张三把李四的车钥匙藏起来了。[s4]他喜欢逗乐。
 - [s5]张三把李四的车钥匙藏起来了。[s6]他爱看电影。
- 意义上是如何关联的？
 - 通过句子的意义表示关联
 - [s1]和[s2]构成合理的篇章：两个句子表示“因果关系”
 - [s3]和[s4]也构成合理篇章：同样表示“因果关系”
 - [s5]和[s6]构成合理篇章吗？
 - 通过句子意义表示的关联称为连贯(**coherence**)
 - 如何解释[s5] 和 [s6]
 - 一种推断：“他希望李四请他看电影”（可能需要更大的上下文）

衔接和连贯

- Cohesion(衔接): 强调其构成成分(主要是词或短语)之间的关联性.
 - 例子:
 - [s1]张三喜欢**骑自行车**上班, [s2]李四通常**步行**去办公室
 - 在词汇层面上相对容易处理
- Coherence(连贯): 强调整体上表达某种意义
 - 例子:
 - [s3] A: 我有两张票, 想请你**今晚看电影**。
 - [s4-1] B:很遗憾, 我**今晚**不能**看电影** (衔接+连贯, 简洁易懂)
 - [s4-2] **B:我还有一大堆的作业没有完成** (连贯, **没有衔接**)
 - [s4-3] **B:我就不客气了** (连贯, **没有衔接**)
 - **[s4-4] B: 武汉又称江城** (不衔接、不连贯)
 - 在处理上相对困难, 不容易切入

衔接的进一步解释

- **Cohesion**: Five cohesive relations (Halliday & Hasan, 1976)
 - Reference (指代)
 - Substitution (替换)
 - Ellipsis (省略)
 - Conjunction (连接)
 - Lexical cohesion (词汇衔接)
- 语篇中为什么会有衔接现象？
 - 追求表达的经济（省略、指代）；
 - 追求表达的变化（指代、替换、词汇衔接）；

词汇衔接

- ▶ 复现关系 (reiteration)
 - ▶ 重复
 - ▶ 同义词和近义词
 - ▶ 上下义词
 - ▶ 泛指词
- ▶ 搭配关系 (collocation)

词汇衔接的例子

- 社交的吃饭种类虽然复杂，性质极其简单。把饭给自己有饭的人吃，那是请饭；自己有饭可吃而去吃人家的饭，那是赏面子。交际的微妙不外乎此。反过来说，把饭给没饭吃的人吃，那是施食，赏面子就一变而成丢脸。这便是慈善救济，算不上交际了。（钱钟书：《吃饭》）。
- 起衔接作用的词
 - 饭
 - 交际（社交）
 - 面子（赏面子、丢脸）
 - 施舍（施食、救济）
 - 复杂（简单）
- 应用：通过衔接关系，可以用于提取文本的关键词

连贯关系 (coherence)

- 语段 (如句子) 之间可能的语义连接关系称为**连贯关系**。
- Hobbs(1979)提出的连贯关系 (设S0和S1为两个相关的句子的**意义**) :
 - 结果关系(Result) : 推测S0所声明的状态或事件 (可能) 导致S1所声明的状态或事件 ;
 - 解释关系(Explanation) : 推测S1所声明的状态或事件 (可能) 导致S0所声明的状态或事件 ;
 - 平行关系(Parallel) : 推测S0所声明的 $P(a_1, a_2, \dots)$ 与S1所声明的 $P(b_1, b_2, \dots)$ 是类似的 ;
 - 细化关系(Elaboration) : 推测S1和S0所声明的是同一命题P ;
 - 时机关系(Occasion) : 推测由S0所声明的状态到S1最终状态的变化 , 或者由S1所声明的状态到S0的最初状态的变化 ;

一个连贯的例子

S1: 张三去银行办理支票.

S2: 然后他乘车到了李四的汽车销售店.

S3: 他想买一部车.

S4: 他的工作单位距公交站较远

S5: 他也想同李四讨论一下他们的垒球协会的事情

Occasion

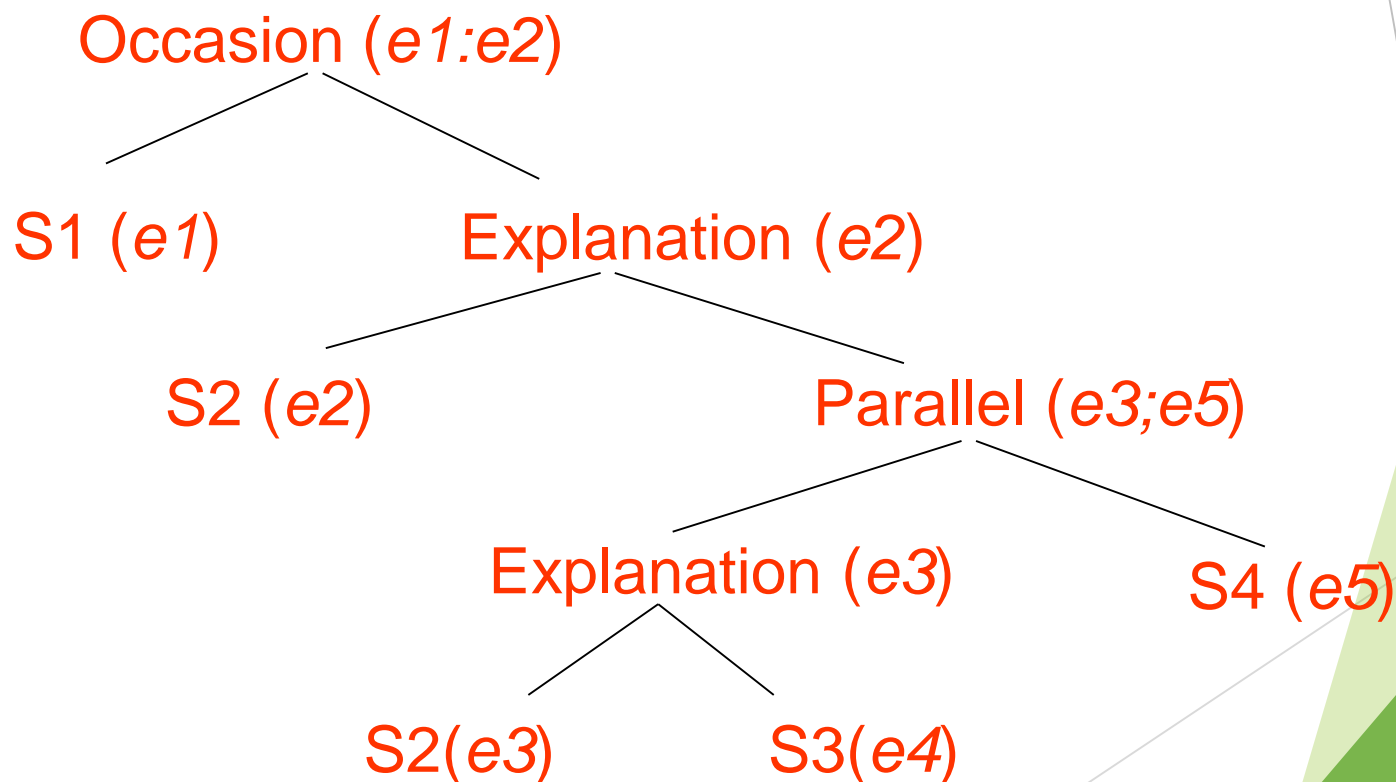
Explanation

Explanation

Parallel

基于连贯的篇章层次结构

- 建立句间语义关系（以前面**5**个句子为例）



修饰结构理论RST

- **修饰结构理论**：认为语篇的构成具有层次结构关系（树形图），通过修饰结构表示语篇结构
- 层次结构关系由**修饰关系**刻画
- 修饰关系是对前面Hobbs连贯关系的细化
 - 共23种关系；
 - 关系的双方：**Nucleus** 与 **Satellite**
 - 具有支配作用：**Nucleus - Satellite**
 - 平等关系：**Nucleus - Nucleus**

RST中的关系

Subject matter (informational)

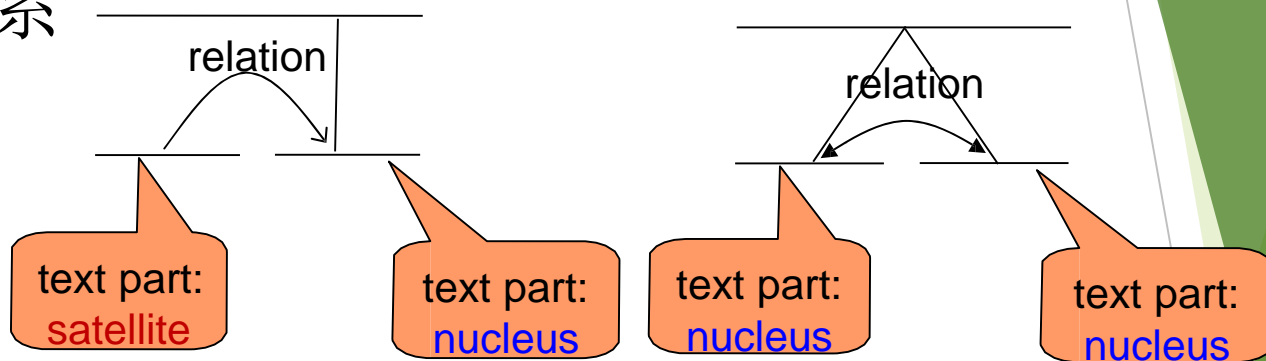
Elaboration
Circumstance
Solutionhood
Volitional Cause
Volitional Result
Non-Volitional Cause
Non-Volitional Result
Purpose
Condition
Otherwise
Interpretation
Evaluation
Restatement
Summary
Sequence
Contrast

Presentation (intentional)

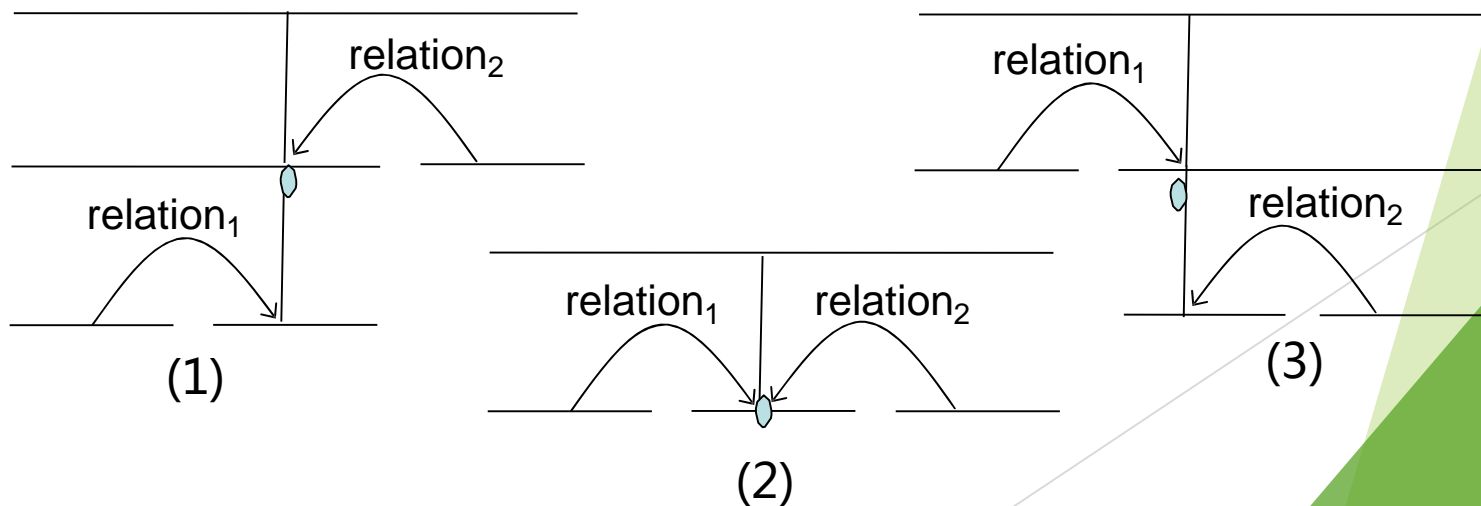
Motivation
Antithesis
Background
Enablement
Evidence
Justify
Concession

基本关系模式

- 二元关系



- 多元素关系



RST- 例子

► 使用RST理论对下面的语篇进行连贯性分析：

1. I love to collect classic automobiles.
2. My favorite car is my 1899 Duryea.
3. However, I prefer to drive my 1999 Toyota.

► 前两句话是什么关系？

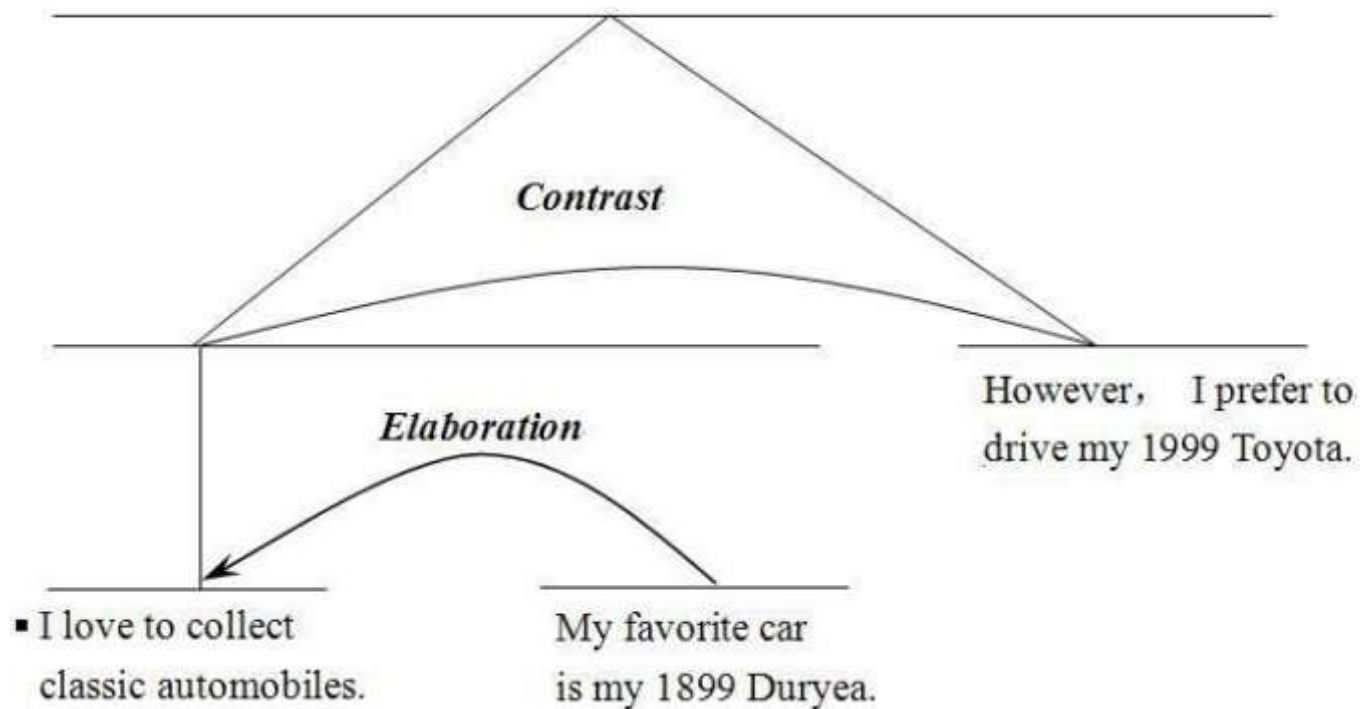
2对1进行了详细说明（Elaboration）

► 3与1和2是什么关系？

*However*明显表示了两者之间的对照（Contrast）关系

RST- 例子

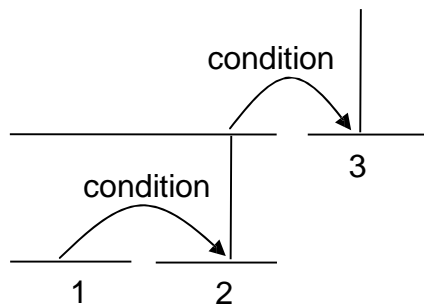
- ▶ 据此建立篇章结构树



多种解释

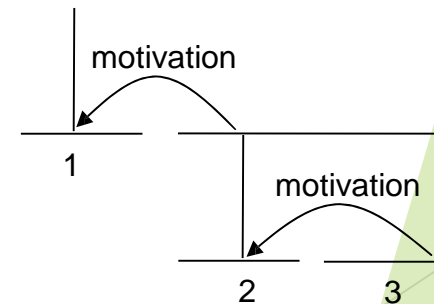
- 1. *Come back at 5:00.*
- 2. *Then we can go to the hardware store before it closes.*
- 3. *This way we can finish the bookshelves tonight.*

Informational level



Condition: The satellite presents a situation which is necessary for the nucleus to obtain.

Intentional level



Motivation: Satellite presents information which should make the reader want to perform the action in the nucleus

RST Discourse Treebank

- ▶ **RST Discourse Treebank** : 基于RST理论对385篇英语篇章进行分析得到的语料库
- ▶ catalog.ldc.upenn.edu/products/LDC2002T07
- ▶ 示例 :

```
( Root (span 1 6)
  ( Nucleus (span 1 3) (rel2par span)
    ( Nucleus (span 1 2) (rel2par span)
      ( Satellite (leaf 1) (rel2par attribution) (text _!Westinghouse Electric Corp. said_!) )
      ( Nucleus (leaf 2) (rel2par span) (text _!it will buy Shaw-Walker Co._!) )
    )
    ( Satellite (leaf 3) (rel2par elaboration-additional) (text _!Terms weren't disclosed.<P>_!) )
  )
  ( Satellite (span 4 6) (rel2par elaboration-additional)
    ( Nucleus (span 4 5) (rel2par Same-Unit)
      ( Nucleus (leaf 4) (rel2par span) (text _!Shaw-Walker,_!) )
      ( Satellite (leaf 5) (rel2par elaboration-additional-e) (text _!based in Muskegon, Mich.,_!) )
    )
    ( Nucleus (leaf 6) (rel2par Same-Unit) (text _!makes metal files and desks._!) )
  )
)
```

目录

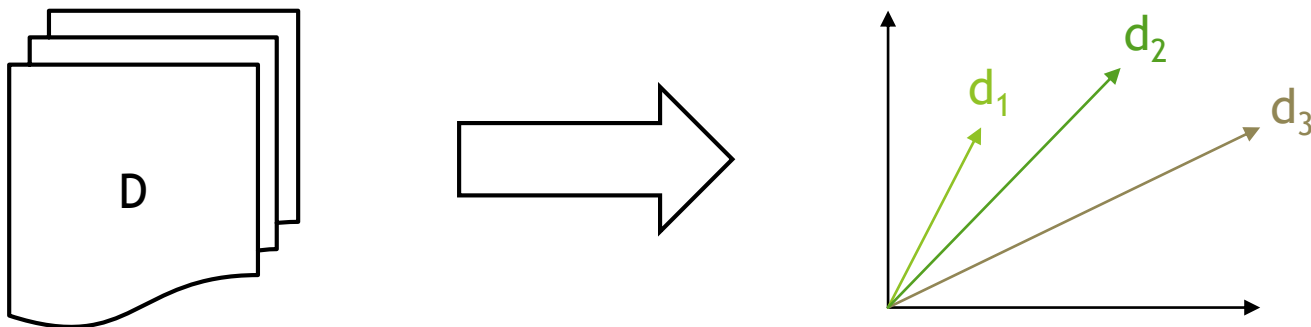
- ▶ 语篇分析简介
- ▶ 指代和指代消解
- ▶ 衔接和连贯
- ▶ **篇章表示和相似度计算**

集合论

- ▶ 文本：词汇的集合
 - ▶ 单词是不是就够了？
 - ▶ 回想一下BLEU: n-gram的集合
- ▶ 问题：

向量空间模型

- **向量空间模型**是一个把文本文件表示为标识符向量的代数模型



- 将两个文本都表示为向量之后，就可以进行相似度的计算

向量的构造方法

- ▶ 将文档表示为如下所示的向量：

$$d_j = (w_{1,j}, w_{2,j}, w_{3,j}, \dots, w_{t,j})$$

- ▶ 向量的每一维都对应于词表中的一个词。

如果某个词出现在了文档中，那它在向量中的值就非零。

- ▶ 这个值有很多计算方法，我们使用词语在文档中出现的次数表示。

向量空间模型的示例

► 对于下面三个文档：

d_1 : "new york times"

d_2 : "new york post"

d_3 : "los angeles times"

► 统计词频：

	new	york	times	post	los	angeles
d_1	1	1	1	0	0	0
d_2	1	1	0	1	0	0
d_3	0	0	1	0	1	1

► 则文档对应的向量为：

d_1 : (1,1,1,0,0,0)

d_2 : (1,1,0,1,0,0)

d_3 : (0,0,1,0,1,1)

相似度计算

► 余弦相似度

$$\text{sim}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{\|\mathbf{d}_i\| \|\mathbf{d}_j\|}$$

► 以之前的文档为例：

$$\text{sim}(d_1, d_2) = \frac{2}{\sqrt{3} \times \sqrt{3}} = 0.67$$

$$\text{sim}(d_1, d_3) = \frac{1}{\sqrt{3} \times \sqrt{3}} = 0.33$$

$$\text{sim}(d_2, d_3) = \frac{0}{\sqrt{3} \times \sqrt{3}} = 0$$

基于句子的语篇表示

- ▶ 分层的语篇表示方法
 1. 通过某种方法得到句子的编码表示
 2. 对句编码进行组合，得到篇章表示
- ▶ 对句子进行编码的方法：向量空间模型、N元语言模型、循环神经网络等等
- ▶ 对句编码进行组合的方法：TextRank、卷积神经网络、循环神经网络等等

篇章（外）的计算

▶ 文本分类

▶ 文本聚类 identify

▶ -----

▶ 自动文摘 manipulate

▶ 文本生成 ***