

自然语言处理

2022年秋季

黄河燕， 鉴萍

北京理工大学 计算机学院

hhy63, pjian@bit.edu.cn

知识体系、问题及方法论

(一) 五大问题、三大范式、语言学基础

黄河燕，鉴萍

北京理工大学 计算机学院

hhy63, pjian@bit.edu.cn

大纲

□ NLP的几个问题

□ NLP的几个研究范式


□ NLP与语言学


□ NLP与知识工程


□ NLP与机器学习


□ NLP与贝叶斯理论和概率图模型

□ NLP与信息论


 Lec5 序列评估与序列标注.pptx

 Lec6 序列结构预测.pptx

 Lec7 自然语言的语义.pptx

 Lec8 序列转换-机器翻译.pptx

 Lec9 序列转换-自然语言生成及相关任务.pptx

 Lec10 文本匹配.pptx

NLP的几个问题

- 分类 (classification)
- 序列评估 (sequence evaluation)
- 序列标注 (sequence labeling)
- 序列结构预测 (sequence structurization)
- 序列转换 (sequence transformation)
- 文本匹配 (text matching)

NLP的几个问题

□ 分类 (classification)

- 输入：一个token或一个句子或一篇文档等
- 输出：一个label
- 典型任务：
 - WSD、实体关系分类、指代消解
 - 文本蕴含、篇章关系分析、复述判断
 - 文本分类、情感分析

Input sentence:

The United States President Biden will meet with Hyundai's CEO Zhong Yishun.

文本分类: {**政治**, 军事, 体育,} 情感分类: {褒义, 贬义, **中性**}

实体关系分类:

Relations: {United States, **Country-President**, Biden} {Hyundai, **Person-Company**, Zhong Yishun}

Input sentence2:

The US President Biden and the CEO of Hyundai will have a meeting.

{**相似**, 不相似}

{**复述**, 非复述}

Input sentence3:

The US President Biden will visit CEO of General Motors Company. → {蕴含, 对立, **中立**}

Input sentence4:

"I voted for **Nader** because **he** was most aligned with **my** values," **she** said. 指代消解

{**I**, **Nader**, **he**, **my**, **she**}

共指**cluster1**, 共指**cluster2**

指代(共指)消解更确切地说是将指示词划分到正确的**cluster**

分类输出可能不是一个**label**, 而是一个概率值(或打分)

NLP的几个问题

□ 序列评估 (sequence evaluation)

- 输入：一个序列
 - 输出：
 - 合法性评估：是/否 (是不是well-formed?)
 - 可能性评估：一个概率值
 - 典型任务：
 - 语言模型(language model)、汉语分词、文本自动校对(proofreading)、语音识别、音字转换(pinyin-to-character conversion)
- 1. 结合成分子时: $P(W) = 0.2$
2. 结合成分时: $P(W) = 0.8$ ✓

NLP的几个问题

□ 序列标注 (sequence labeling)

- 输入：一个序列
- 输出：一个tag序列，每个token标注一个tag
- 典型任务：
 - 音字转换：拼音序列 → 字序列
 - 词性标注：词序列 → 词性序列
 - 命名实体识别：词/字序列 → 实体标注序列
 - 组块分析：词序列 → 短语标注序列

湖北 武汉 正在 上演 樱花 主题 灯光秀

B_LC E_LC O O O O O

NLP的几个问题

□ 序列结构预测 (sequence structurization)

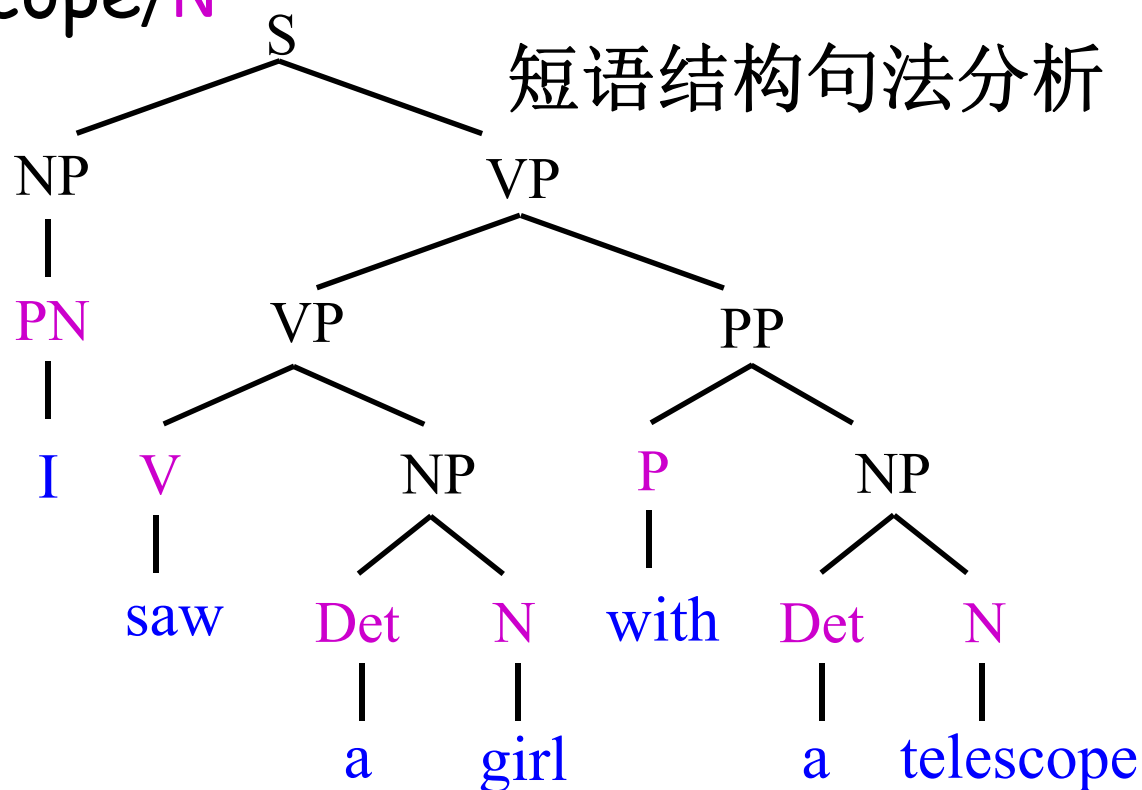
- 输入：一个序列
- 输出：一个描述element间关系的结构
- 典型任务：
 - 成分(短语结构句法)分析：词序列 → 短语结构树
 - 依存句法分析：词序列 → 依存结构树
 - 语义分析：词序列 → 逻辑形式
 - 语义依存分析：词序列 → 语义依存结构树
 - 篇章结构分析：篇章单元序列 → 篇章结构

NLP的几个问题

□ 序列结构预测 (sequence structurization)

➤ 输入: I/PN saw/V a/Det girl/N with/P
a/Det telescope/N

输出:



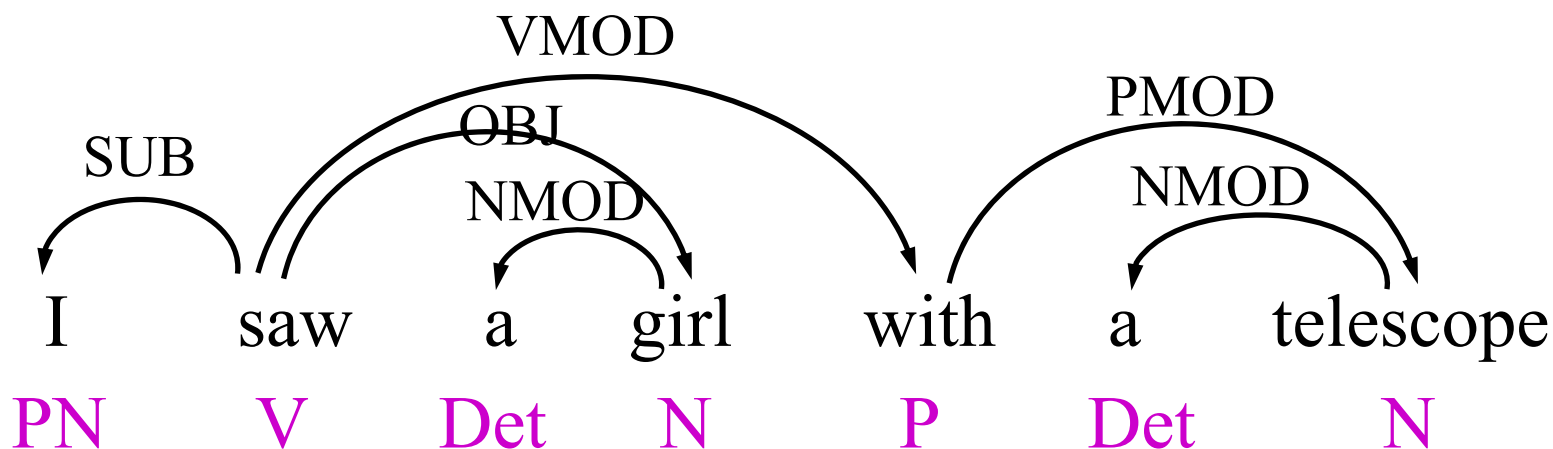
NLP的几个问题

□ 序列结构预测 (sequence structurization)

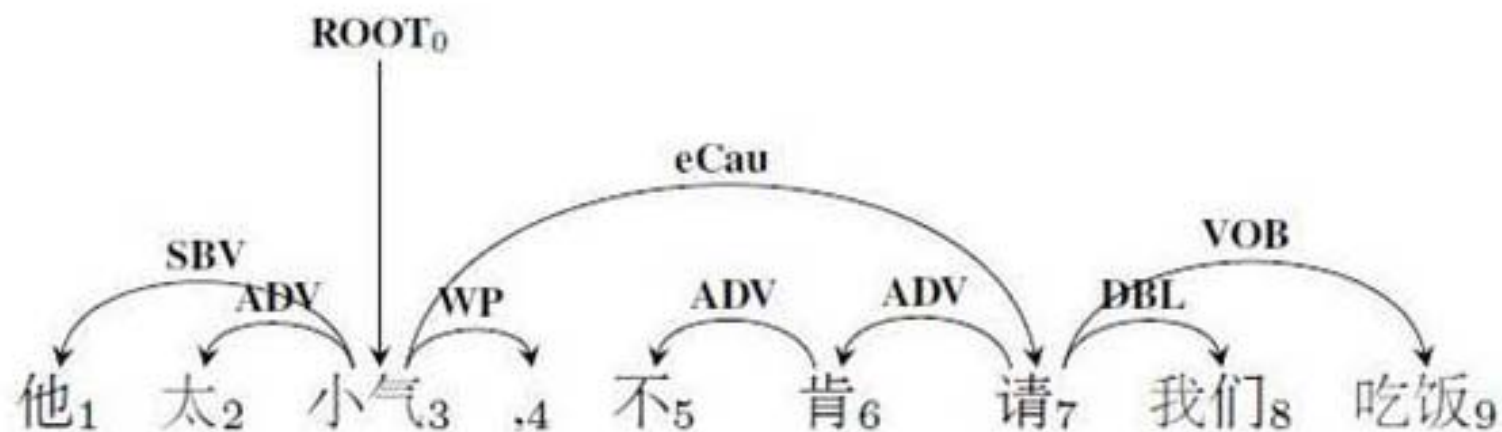
➤ 输入: I/PN saw/V a/Det girl/N with/P
a/Det telescope/N

输出:

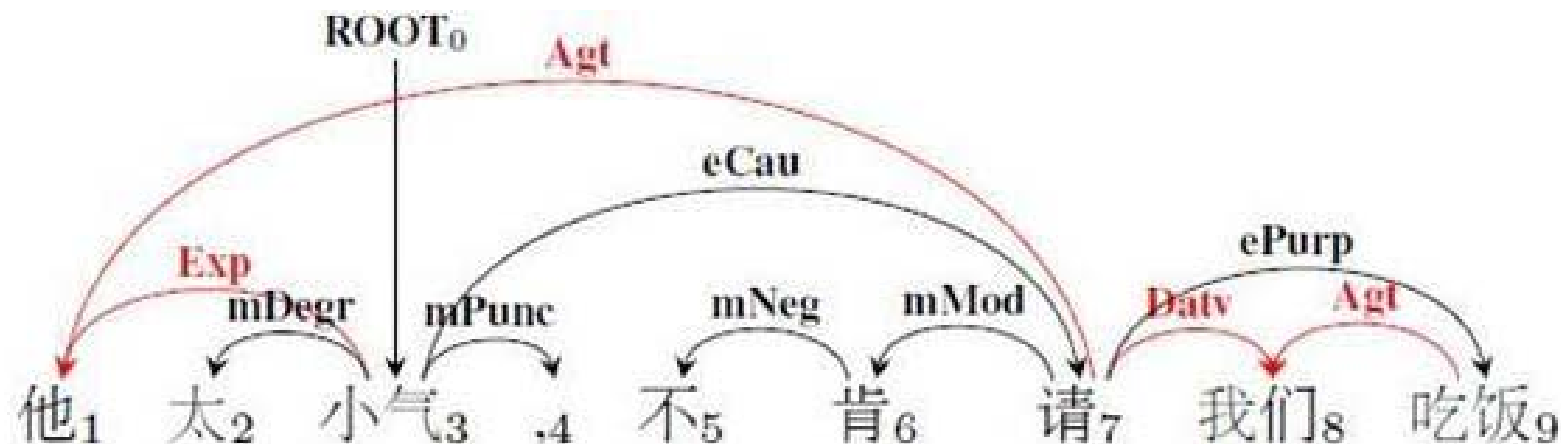
依存句法分析



NLP的几个问题

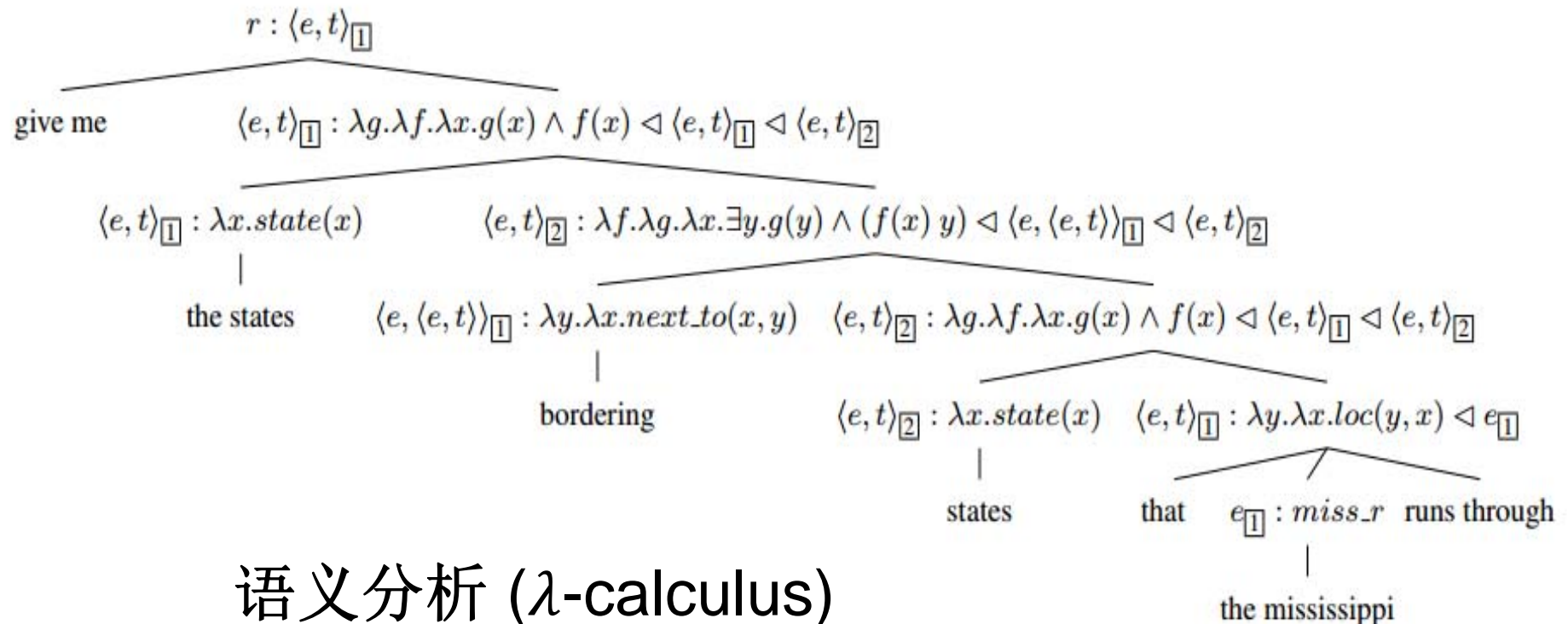


句法依存树 **vs.** 语义依存树



NLP的几个问题

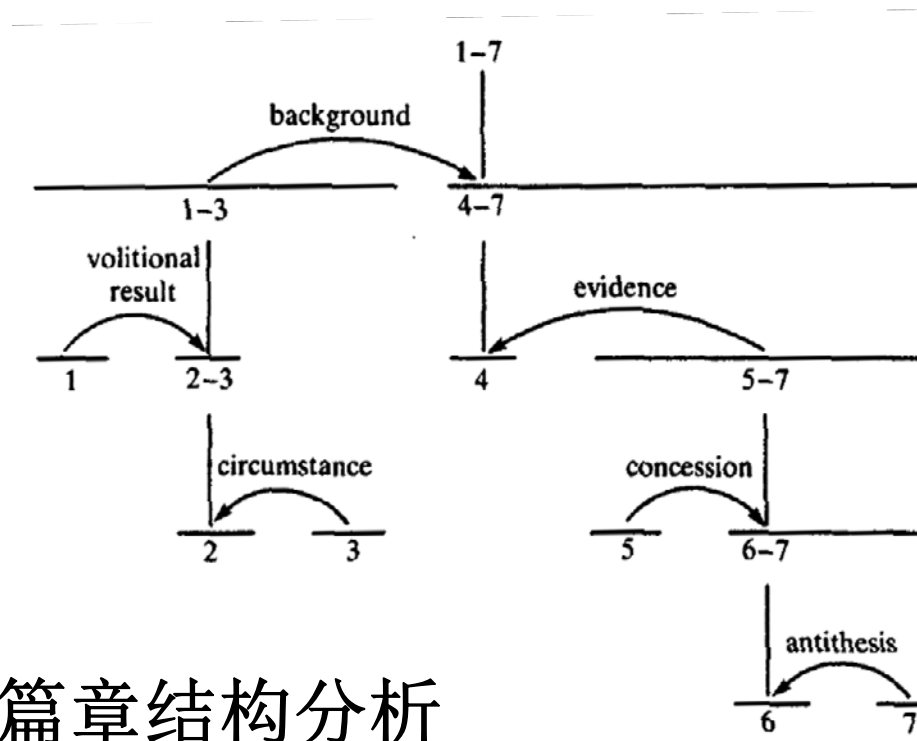
□ 序列结构预测 (sequence structurization)



语义分析 (λ -calculus)

NLP的几个问题

□ 序列结构预测 (sequence structurization)



NLP的几个问题

□ 序列转换 (sequence transformation)

- 输入：一个序列 (自然语言句子)
- 输出：另一个序列 (自然语言句子)
- 典型任务：

- 机器翻译、复述、对话、摘要、续写

- 复述：

输入：怎样做红烧肉？

输出：红烧肉的做法

NLP的几个问题

□ 文本匹配 (text matching)

- 输入：一个query，一个doc(或者KB)
- 输出：一个label——是否匹配
一个answer——匹配出的答案
- 典型任务：
 - 信息检索，相似度计算，复述检测，文本蕴含，问答(基于知识库KB的问答和基于文档的问答)
(文本匹配范畴很广，它也可以是分类或生成问题)

段落

工商协进会报告，12月消费者信心上升到78.1，明显高于11月的72。另据《华尔街日报》报道，2013年是1995年以来美国股市表现最好的一年。这一年里，投资美国股市的明智做法是追着“傻钱”跑。所谓的“傻钱”策略，其实就是买入并持有美国股票这样的普通组合。这个策略要比对冲基金和其它专业投资者使用的更为复杂的投资方法效果好得多。

问题1：什么是傻钱策略？

答案：买入并持有美国股票这样的普通组合

问题2：12月的消费者信心指数是多少？

答案：78.1

问题3：消费者信心指数由什么机构发布？

答案：工商协进会

基于文档的问答(阅读理解)

Passage: Alice's mother died when she was five. Although her brothers and sisters were loving and caring, their love couldn't take the place of a mother's. In 1925 Alice became my mother and told me that her family couldn't even afford her a doll. One afternoon in December 1982, when I was getting ready for Christmas, I suddenly decided to buy two dolls, one for my five-year-old daughter, Katie, and one for my old mother. Things went smoothly when a friend told me that his dad, who played Santa Claus in my area, would be willing to make a visit on Christmas morning to our home with the gifts! Knowing that my parents would also come to my house, I began to get ready for the most memorable day of my mother's life. Christmas Day arrived and so did Santa Claus at the planned time...

Question: Why didn't Alice expect there was also a gift for her?

- A. The gifts from Santa Claus were usually for children.
 - B. The gift was forgotten many years ago.
 - C. The gift for her was bought by accident on the way.
 - D. The gifts for Katie were enough to share with her
-

NLP的几个问题

□ 不同问题间可以相互转换

- 序列评估 → 序列标注
- 序列标注 → 分类
- 序列结构预测 → 分类
- 文本匹配 → 分类
- 序列标注 → 序列转换

你能活到付款吗？

翻译

————→ 你能货到付款吗？

大纲

- NLP的几个问题
- NLP的几个研究范式
- NLP与语言学
- NLP与知识工程
- NLP与机器学习
- NLP与贝叶斯理论和概率图模型
- NLP与信息论

NLP的研究范式

符号主义 (规则系统) \Rightarrow 概率统计方法 (概率图模型) \Rightarrow 深度学习方法 (深度神经网络)

语言学家, 逻辑学家
数理逻辑

计算机科学家
概率论, 数理统计

计算机科学家
线性代数, 微积分

□ 从AI的三大流派来看

- 符号主义(symbolicism)
- 连接主义(connectionism)
- 行为主义(actionism)

NLP的研究范式

□ 规则系统——规则举例

(V **ROOT** ?r **SUBCAT** ?s **VFORM** pres **AGR** 3s) →

(V **ROOT** ?r **SUBCAT** ?s **VFORM** base **IRREG - PRES**—) + S

(V **ROOT** ?r **SUBCAT** ?s **VFORM** pres **AGR** {1s 2s 1p 2p 3p}) →

(V **ROOT** ?r **SUBCAT** ?s **VFORM** base **IRREG - PRES**—)

词法规则，用于形态变化

NLP的研究范式

□ 规则系统

if (termination of **word** is *ied*) *then*

Copy **word** to **word1**

Eliminate *ied* of **word1** and add ‘y’

if (**word1** is in the lexicon) *then*

Add attributes **PAST** and **VEN** to **word**

else if (termination of **word** is *ed*) *then*

Copy **word** to **word1**

Eliminate *ed* of **word1**

if (**word1** is in the lexicon) *then*

Add attributes **PAST** and **VEN** to **word**

词法分析程序，用于形态分析

NLP的研究范式

□ 规则系统——规则举例

Change tag *a* to tag *b* when:

- The preceding (following) word is tagged *z*.
- The word two before (after) is tagged *z*.
- One of the two preceding (following) words is tagged *z*.
- ...

词性标注系统里的修正规则

NLP的研究范式

□ 规则系统——规则举例

- 基于规则的机器翻译系统中的一个句法分析规则：

LEX=站, (base0), **RULE**: (1111; 113; 1113) + ^ (CSUBCAT.vv2; CSUBCAT.vv1; CSUBCAT.vv3; CSUBCAT.vv4) \Rightarrow 1^.改.
^ \downarrow (SYNRELA.sub)

当前词前面一个词的语义分类码为**1111**或**113**或**1113**的词(这里一般指有动物性的名词，即能发出动作的)，当前词的下位词性为**vv2**或**vv1**或**vv3**或**vv4**的词，则当前词左边的第一个词(在这里是**女孩**)归为当前词(这里是**站**)的孩子，并且句法关系是**sub**，主语。

NLP的研究范式

➤ 一个超级简单的句子情感判别器

词的情感打分加和平均，得到句子的情感打分

	A	B
1	word	strength
2	百分之百	3
3	倍加	3
4	备至	3
5	不得了	3

full_pos_dict_sougou		full_neg_dict_sougou	
1	清莹	1	脏乱
2	轻倩	2	糟报
3	晴丽	3	早衰
4	求索	4	责备
5	热潮	5	贼眼

查情感词典

这 汽水 太 凉 。

NLP的研究范式

- 但是，自然语言这么灵活，新词层出不穷，规则能cover所有的语言现象吗？规则的冲突和冗余怎么办？

社恐，社死？社牛？

躺平，摆烂？

吃席？



绵阳高温山洞里长满了避暑的居民

近日，受高温天气影响，四川绵阳群众扎堆山洞内避暑。山洞里面有人在打牌，还有小孩和宠物。村民称，洞内只有十几度，还需要穿两件衣服。[详细 >](#)

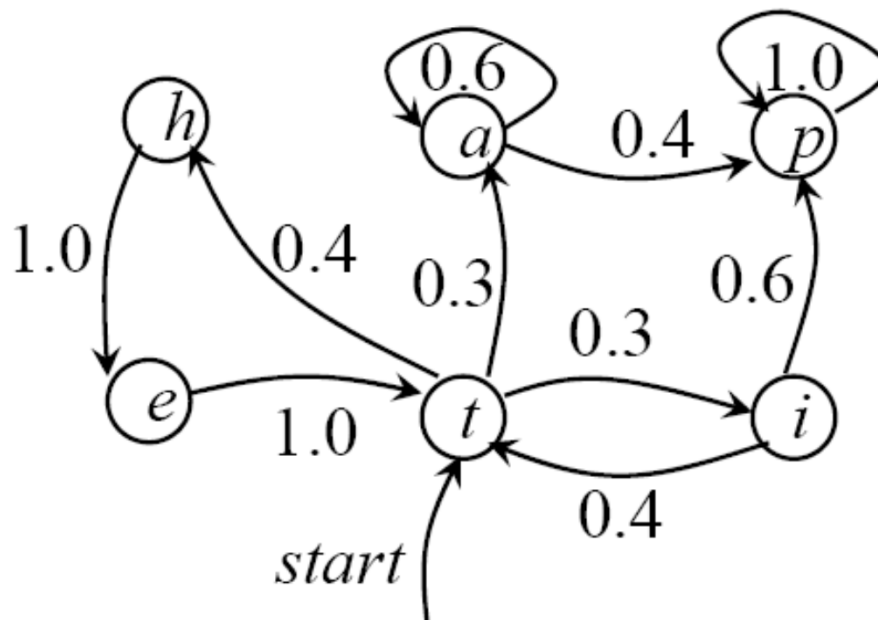
椒点资讯 5小时前

NLP的研究范式

□ 推理只是人类智能的一种，人们还希望用类比的方法，逐步积累经验来认知世界

□ 概率统计模型

从演绎到归纳



$$\begin{aligned} P(\mathbf{t-i-p}) &= P(q_1 = \mathbf{t}) \times P(q_2 = \mathbf{i} | q_1 = \mathbf{t}) \times P(q_3 = \mathbf{p} | q_2 = \mathbf{i}) \\ &= 1.0 \times 0.3 \times 0.6 = 0.18 \end{aligned}$$

已知统计
数据:

First tag	Second tag					
	AT	BEZ	IN	NN	VB	PERIOD
AT	0	0	0	48636	0	19
BEZ	1973	0	426	187	0	38
IN	43322	0	1325	17314	0	185
NN	1067	3720	42470	11773	614	21392
VB	6072	42	4758	1476	129	1522
PERIOD	8016	75	4656	1329	954	0

$P(\text{AT NN BEZ IN AT NN} \mid \text{The bear is on the move})$

$P(\text{AT NN BEZ IN AT VB} \mid \text{The bear is on the move})$

哪个更有可
能是这句话
的词性序列?

	AT	BEZ	IN	NN	VB	PERIOD
<i>bear</i>	0	0	0	10	43	0
<i>is</i>	0	10065	0	0	0	0
<i>move</i>	0	0	0	36	133	0
<i>on</i>	0	0	5484	0	0	0
<i>president</i>	0	0	0	382	0	0
<i>progress</i>	0	0	0	108	4	0
<i>the</i>	69016	0	0	0	0	0
	0	0	0	0	0	48809

NLP的研究范式

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value									

文档中词的分布可以决定文档的主题分布

NLP的研究范式

基于噪声信道的统计翻译模型

$$\begin{aligned} e^* &= \arg \max_e P(e|f) = \arg \max_e \frac{P(f|e)P(e)}{P(f)} \\ &= \arg \max_e P(f|e)P(e) \end{aligned}$$

翻译模型

语言模型

人工评价指标:

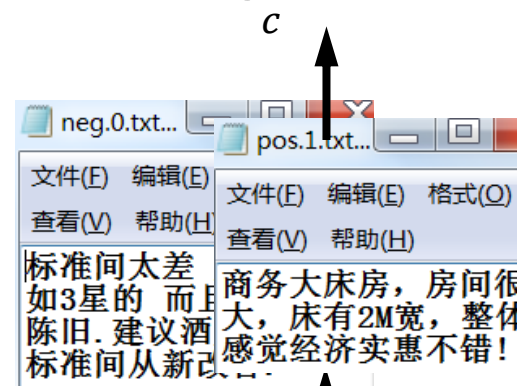
adequacy

fluency

NLP的研究范式

➤ 一个基于naïve bayes的情感分析器

$$C^* = \operatorname{argmax}_C P(C_i|W)$$



标注语料中查 $P(w_k|C)$

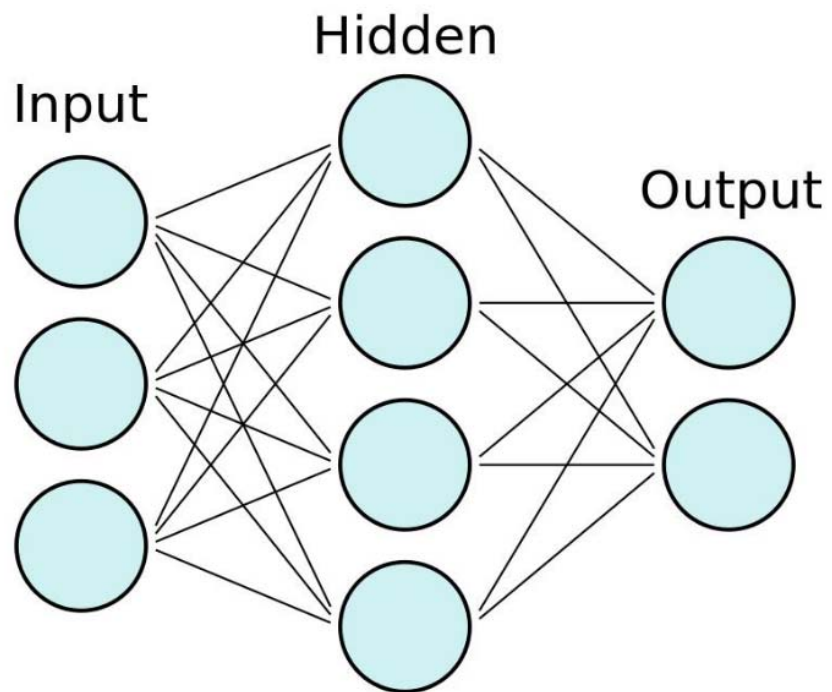
$$P(C|W) = \frac{P(W|C)P(C)}{P(W)}$$
$$= \frac{\prod_k P(w_k|C)P(C)}{P(W)}$$

Naïve Bayes

这 汽水 太 凉 。

NLP的研究范式

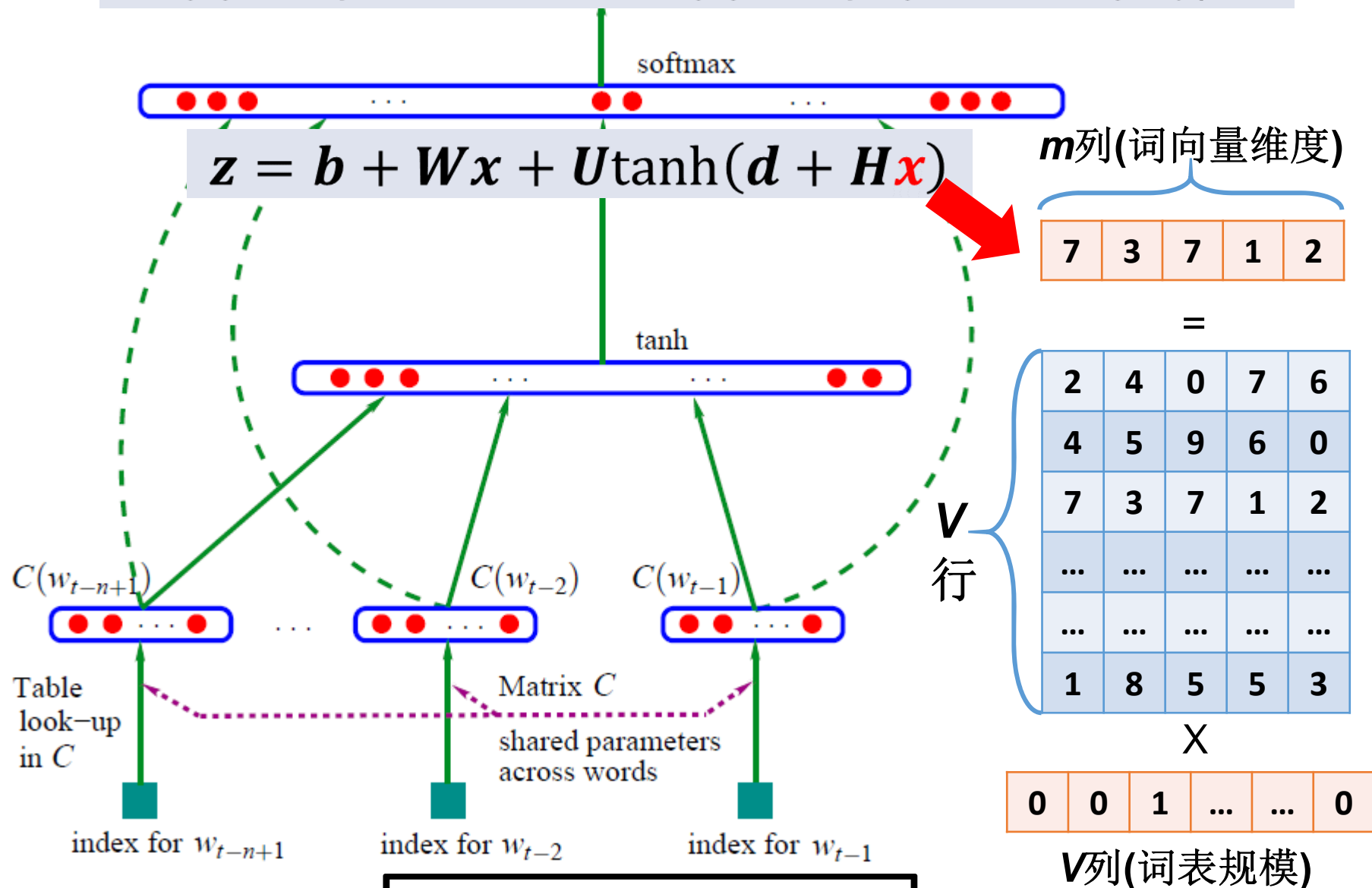
▣ 深度神经网络模型



$$h = g(WX + b)$$

$$y = g(Wh + b)$$

$$P(y_t = i | \text{Context}) = P(y_t = i | w_{t-1}, \dots, w_{t-n+1})$$



$C^{V \times m}$ is trainable

NLP的研究范式

□ 概率统计模型中的词

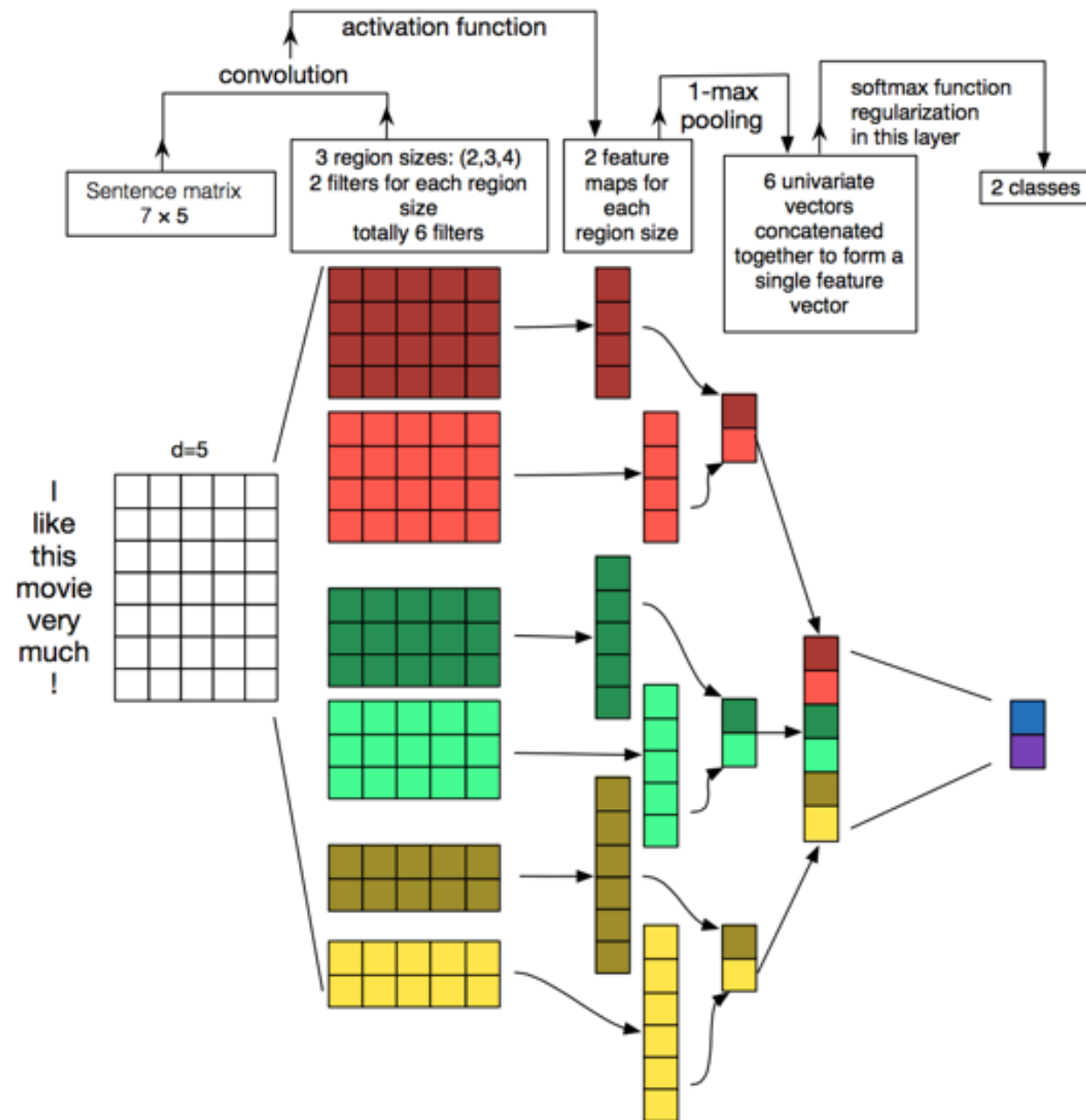
cat: [1 0 ... 0 0 0 0 0 0...]
dog: [0 0 ... 0 0 1 0 0 0...]
walking: [0 1 ... 0 0 0 0 0 0...]
running: [0 0 ... 0 0 0 0 0 1...]

维数是词
表长度
One hot

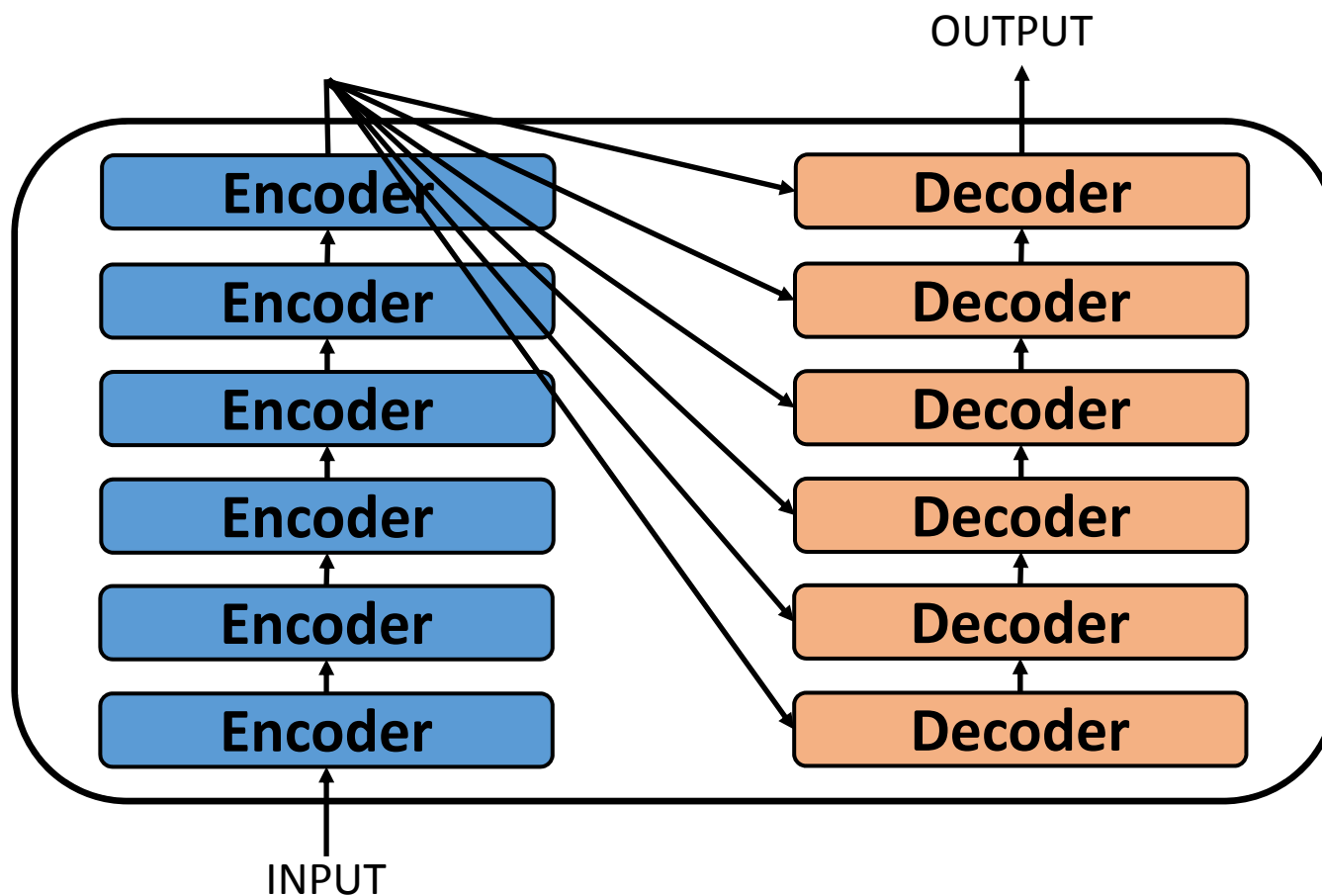
□ 深度神经网络模型中的词

cat: [3.7 -0.17 ... 1.0 0.03 0 -2.5 1.98]
dog: [3.4 -1.2 ... 0.9 -0.07 1.3 0 2.0003]
...

200~500维的稠密向量



NLP的研究范式



NLP的研究范式

□ 深度神经网络模型

- 模型不再可解释
- 看不到推理的踪影
- 效果几乎全靠调参
- 但就是效果好

NLP的研究范式

□ 一个基于逻辑回归的

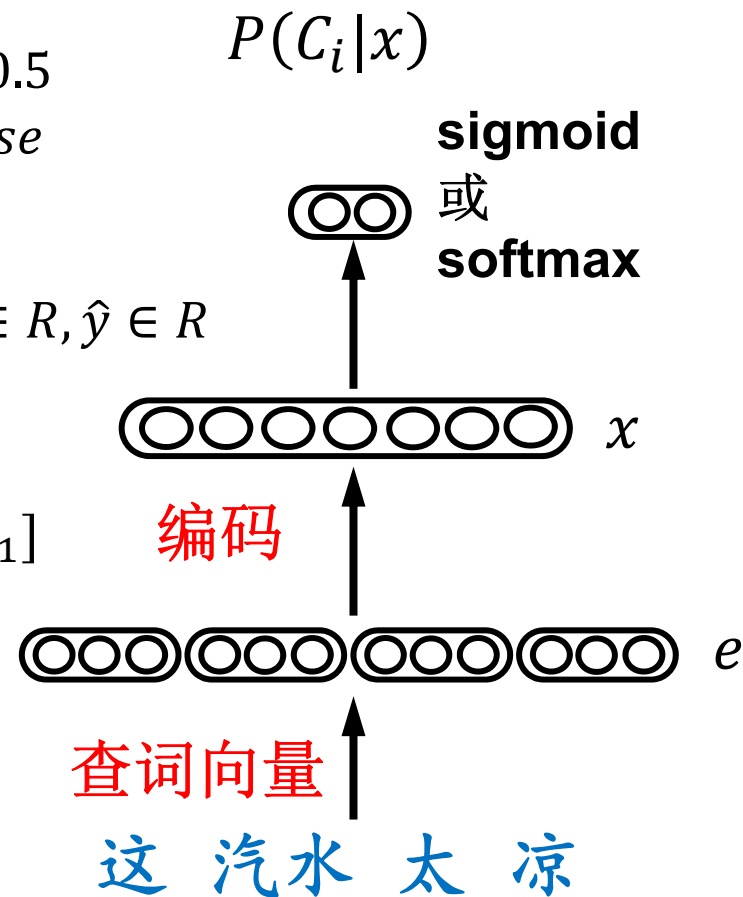
情感分析器

$$\text{类别} = \begin{cases} 0, & \hat{y} < 0.5 \\ 1, & \text{else} \end{cases}$$

$$\hat{y} = \text{sigmoid} \left(b + \sum_{i=0}^{m-1} W_i * x_i \right), \quad W \in R^m, b \in R, \hat{y} \in R$$

$$x = \sum_{i=1}^4 e_i, \quad x \in R^m, x = [x_0, x_1, \dots, x_{m-1}]$$

$$\begin{matrix} e_1 & e_2 & e_3 & e_4 \\ e_i \in R^m, e_i = [e_{i0}, e_{i1}, \dots, e_{i(m-1)}] \end{matrix}$$



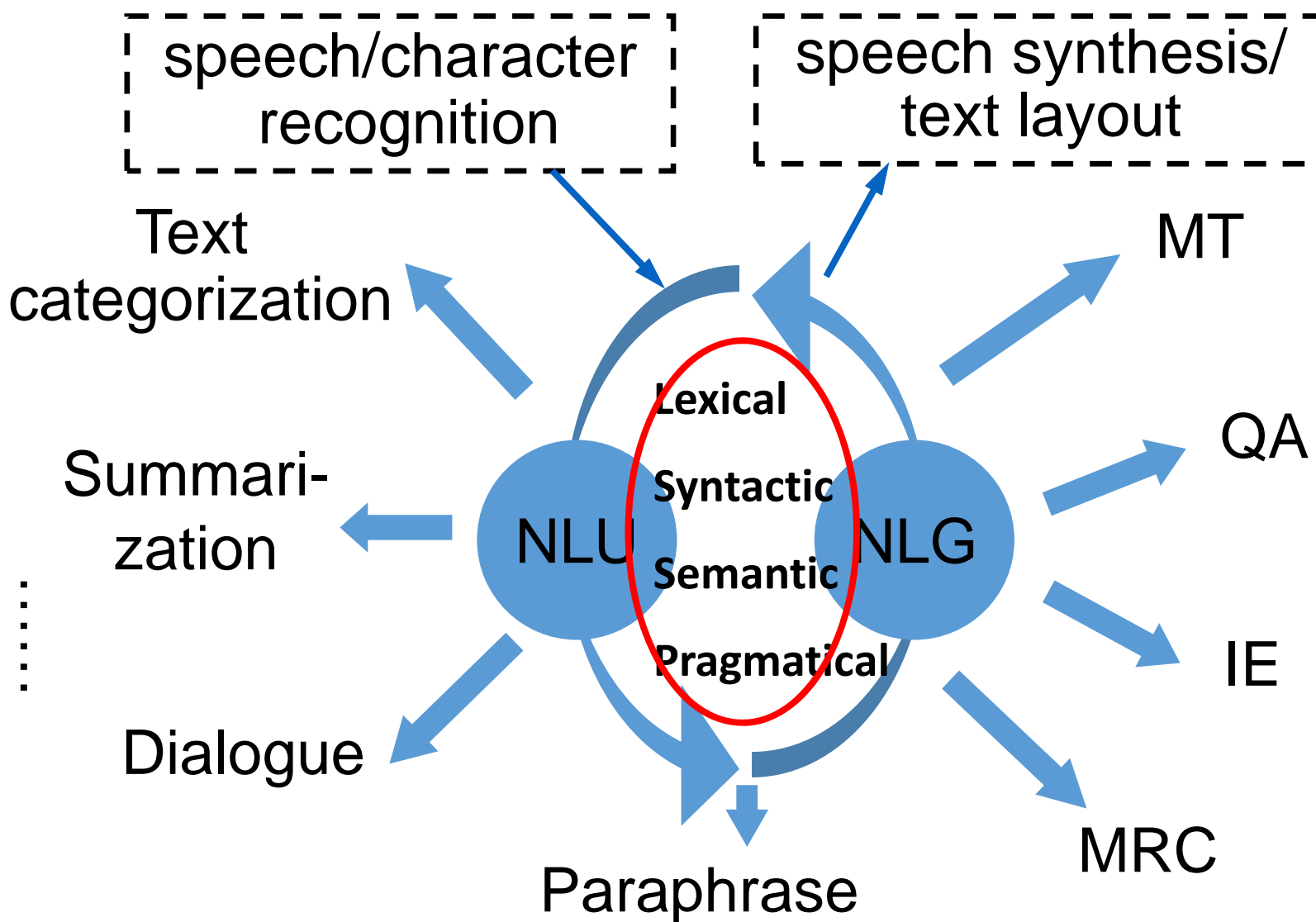
大纲

- NLP的几个问题
- NLP的几个研究范式
- NLP与语言学
- NLP与知识工程
- NLP与机器学习
- NLP与贝叶斯理论和概率图模型
- NLP与信息论



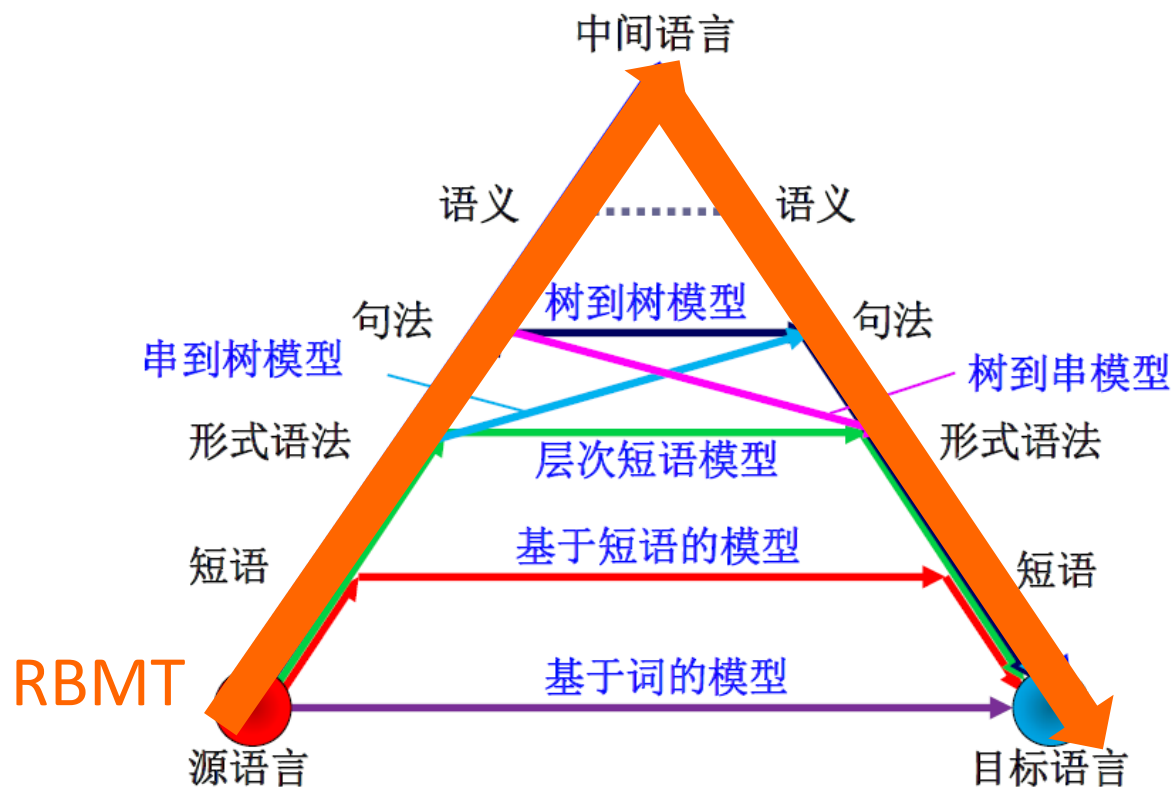
Fred Jelinek 1980s: 我
每开除一名语言学家，
我的语音识别系统错误
率就降低一个百分点。

回顾NLP主要任务



回顾NLP发展历程

□ MT的失败引发NLU的研究



NLP与语言学

□ 自然语言处理的研究遵循语言学对语言的处理层次

- 音位层——音韵分析
- 词法层——词法分析
- 句法层——句法分析
- 语义层——语义分析
- 语用层——语用分析

NLP与语言学

- 什么是语言

- 基本概念

- 词法基础

- 句法基础

- 语义基础

- 语用基础

语言学基础

□ 什么是语言(language) ?

➤ 小朋友蔬菜卡片认知:

我拿出一张说这是南瓜，她说这不是蓝瓜是黄瓜！我拿出另一张卡片说这才是黄瓜，她说这是绿瓜。

饼饼最近每天从幼儿园出来，手心里都攥着个东西，有时候是小石子，有时候是蔫了的花，昨天回家路上，突然就蹲下来画画。

我：哎？你从幼儿园咪的粉笔吗？

她：这怎么是粉笔？！这不是橙笔嘛？！

我：这种笔就叫粉笔啊！

她把我拉到这辆车前面，恨铁不成钢地教我：下面这才是粉！上面这个不是橙嘛？！

我：它是橙的没错，但它画着不是掉粉么！所以就叫粉笔！

她：掉粉是什么？！

我：掉粉... ..就是我一发文章就有人取关。



来自某公
众号

语言学基础

□ 什么是语言(language) ?

➤ 乞丐：“我什么也看不见！”

让·彼浩勒：“春天到了，可是我什么也看不见！”

语言学基础



这个星球（也就是地球）上的人没有记忆传承，关于文明的记忆只能通过教师和书本得以传递，没想到这种效率低得惊人的传承方式，还能够创造出这样发达的文明。



语言学基础

□ 什么是语言(language) ?

在记忆传承得以实现之前，我们人还是得通过非常原始的方法来学习，也就是把具体的事物在头脑中建立表征，然后用符号（比如语言）进行抽象，实现人与人之间的交流，再给抽象的符号赋予形式上的意义，也就是产生字形，实现跨越时间和空间的交流。

——语言是一种符号系统

什么是语言

不论一只狗叫得多么卖力，它也无法对你说明它的父母贫穷却又诚实。

——Bertrand Russell

- 语言的特性

Arbitrariness(任意性) 语言是一种交流工具

Duality(二层性) 人类的语言是离散的可组合

Creativity(创造性) 源于二层性和递归性

Displacement(移位性) 能够谈及已不存在或还未出现的事物，赋予概括与抽象的能力

NLP与语言学

- 什么是语言

- 基本概念

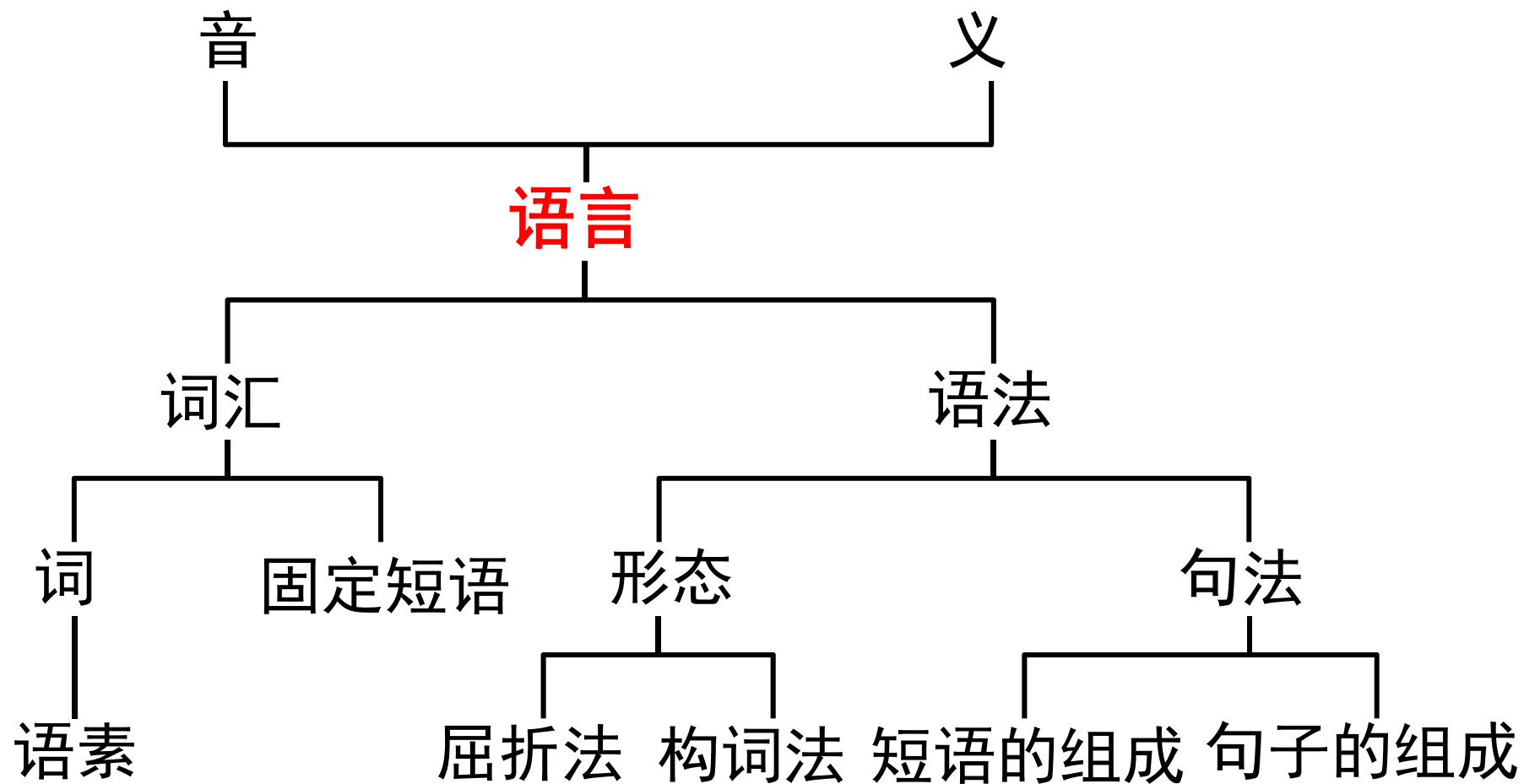
- 词法基础

- 句法基础

- 语义基础

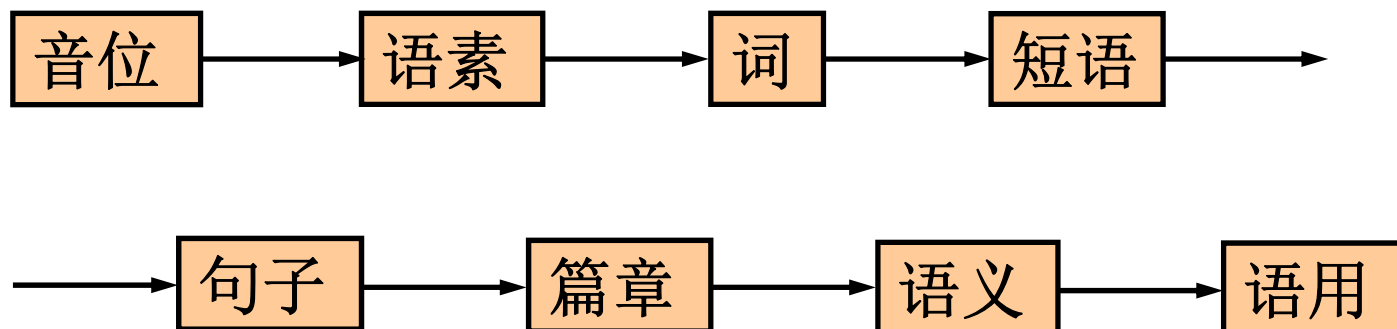
- 语用基础

基本概念



基本概念

□ 语言学这样来研究语言



语言学的著作通常是这样子：

《现代汉语把字句的多角度探究》

《把字句及其英文表达研究》

基本概念

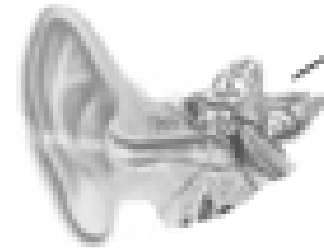
□ 语音学(phonetics)



发声
articulatory



声学
acoustic



听觉
auditory

□ 音韵学(phonology)

- ## 音位 /p/的两种发音:

peak [pi:k], **speak** [spi:k]

Rule: /p/ → [p] / [s] _____
[ph] others

p在s后面不送气，读p（也就是b的发音）
在其它后面送气，读ph（也就是p的发音）

- 56

基本概念

□ 音韵学(phonology)

- 超音段问题(suprasegmental): 研究构成句子时发音的规律, 如音节(syllable)、韵律(prosody)、语调(intonation)、音色(tone)、重音(accent)等

The boys who are ↗ ill can't come.

The boys, who are ↘ ill, can't come.

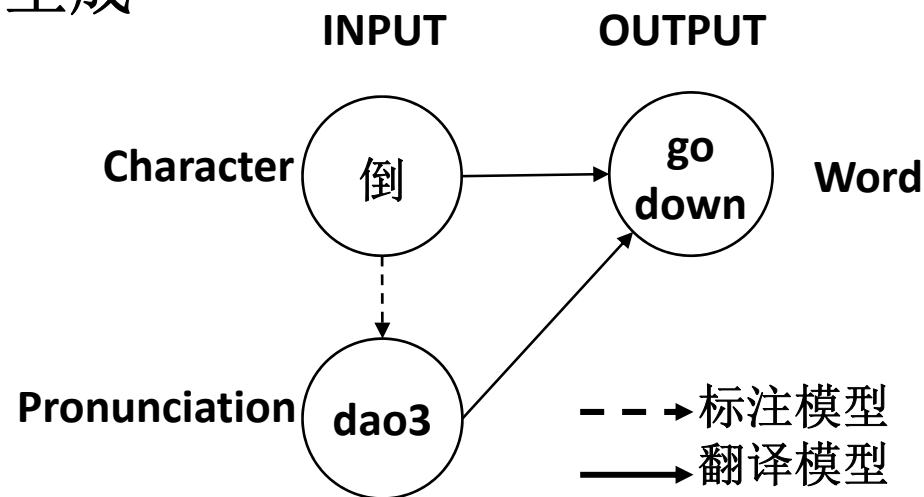
上句上升语调, 限定的定语从句

下句下降语调, 非限定定语从句

基本概念

- 韵律特征在统计机器翻译中应用举例
 - 在汉英翻译中，汉语的韵律特征可以作为输入因子，辅助汉语分析。在英汉翻译中，这些韵律特征作为输出因子，辅助汉语生成

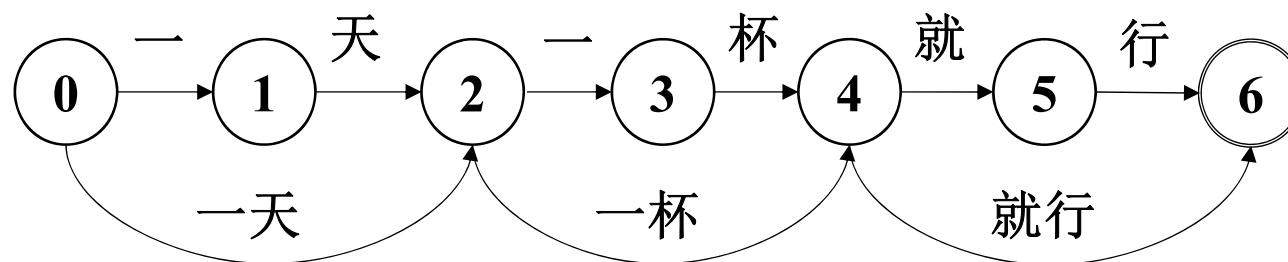
源语言	上 周 日 台 风 ， 倒 dao3 了一 颗 木 瓜 树
参考译文	A papaya tree went down in the typhoon last Sunday
基线系统输出	There is a typhoon on Sunday, pour a papaya tree



引入韵律特征的因子化翻译模型

基本概念

- 韵律特征在统计机器翻译中应用举例
 - 将汉语的韵律边界与字边界共同构建词格，以扩充解码路径，从而增加优质候选，提高译文的质量



基本概念

□ 词(word)：语言中最小的可以独立运用的音义结合的单位，是自然语言处理的基本单位

➤ care, careless, careful

书包、非常、崎岖、小李

□ 语素(morpheme)：语言中最小的音义结合体，是词的构成要素

➤ work|er|s, care|ful

书|包|、崎|岖|、小|李|

基本概念

□ 语素的分类

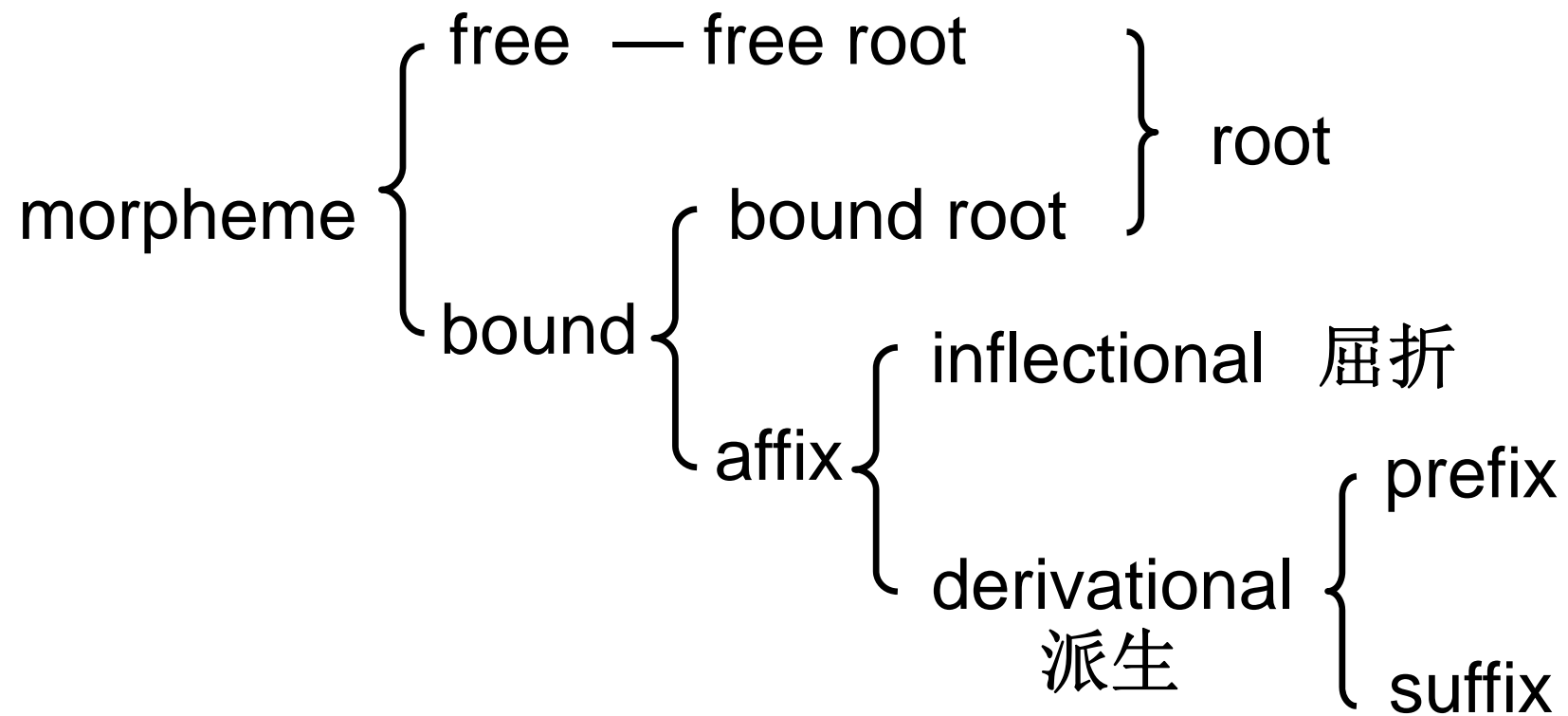
- 自由语素(free): 可以单独成词

- water, child hood, 书包, 我们

- 黏着语素(bound)

- care ful, child hood, 我们, aerial

基本概念



基本概念

□ 词 vs. 语素

1: 单音节词，
是单音节语素，
也是单语素词

2: 双音节词，拆开每个字都失去了意义，是双音节语素，是单语素词

1. 天，地，人，马，跑

2. 徘徊，玫瑰，崎岖，仿佛

3. 葡萄，扑克，迪斯科，幽默

4. 伟大，我们，牙齿，毛巾

5. 白菜，短语，马车，伤心

3: 外来词，
仍然是一个
语素；以上
三组都是单
纯词。

4: 双音节词，双语素词，可以拆分为两个有意义的单位。“我”可以单独使用，既是语素又是词；“们”不可单独使用，仅是一个语素。

5: 和4类似，是双语素组成的合成词，不同的是它的两个组成语素都可以是词。

基本概念

□ 短语(phrase)：指语言结构中通常包含一个以上的词的成分，缺乏典型的主谓结构

- 结构上：陈述，从属，修饰，同位，互补
- 功能上：名词短语，动词短语，介词短语

- 词 vs. 短语

- 这个中药叫红花，红花还得绿叶扶
- 中国科学技术大学，中科大

- 短语在词之上句子之下，可以嵌套即具有层次，是句子结构的直观分解形式

(短语结构句法分析)

基本概念

- 自然语言处理中的“短语”有更大范畴
- 统计机器翻译(SMT)中的短语表：

michael – michael

michael assumes – michael geht davon aus ; michael geht davon aus ,

michael assumes that – michael geht davon aus , dass

michael assumes that he – michael geht davon aus , dass er

michael assumes that he will stay in the house

– michael geht davon aus , dass er im haus bleibt

assumes – geht davon aus ; geht davon aus ,

assumes that – geht davon aus , dass

assumes that he – geht davon aus , dass er

实践证明，统计出来的往往更好

基本概念

□ 句子

- 由语法规则组织，成分相互联系
- 表达相对完整的意思
- 具有一定的语调和情态

以形统意

✓ 英语：一个完整的句法结构是一个句子

✓ 汉语：一个完整的意思(thought)是一个句子

以意统形

He is working hard. 不能写成 * He is work hard.

He is working hard, (必须有连接词如and) he is a good student.

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。

基本概念

接着，他继续设想，鸡又生鸡，用鸡卖钱，钱买母牛，母牛繁殖，卖牛得钱，用钱放债，这么一连串的发财计划，当然也不能算是生产计划。

——《一个鸡蛋的家当》

Translation: He went on indulging in wishful thinking. Chickens would buy mom chickens. Selling them would bring him money. With this he could buy cows. The cows would become mom and selling oxen would make more money for him. With the money, he could become a money lender. Such a succession of steps of getting rich, of course had nothing at all to do with production.

基本概念

黛玉自在枕上感念宝钗……

又听见窗外竹梢蕉叶之上，雨声淅沥，清寒透幕，不觉又滴下泪来。

Translation : As she lay there alone, Dai-yu's thoughts turned to Bao-chai...

Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she had begun to cry.

基本概念

□ 语法

- 广义：写作的艺术
- 狭义：结构的组织：
 - 形态学(词法, **morphology**)：语素如何构成词
 - ◆ book → books
 - ◆ classroom, 老乡, 政协
 - 句法(**syntax**)：词(短语)如何构成句子

NLP与语言学

- 什么是语言
- 基本概念
- 词法基础
- 句法基础
- 语义基础
- 语用基础

词法

- 词的构成
- 词法规则
- 词法分析

词法

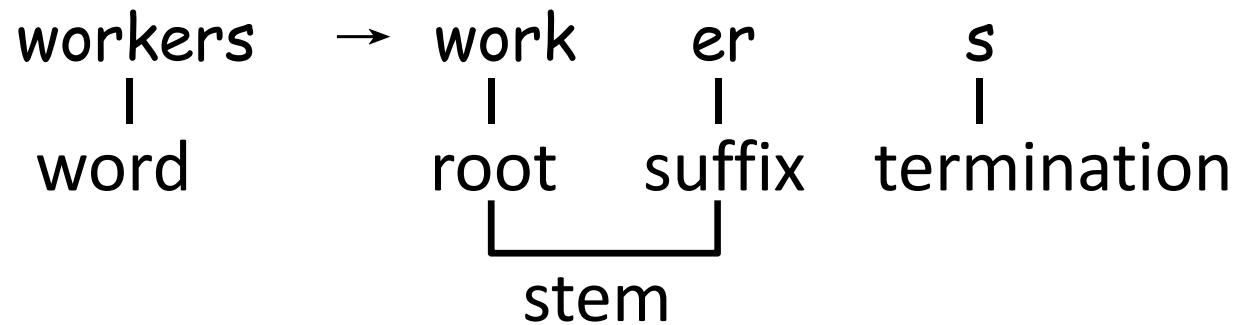
□ 词的构成

word = stem + termination

屈折

stem = prefix + root + suffix

派生



stem: 词干

root: 词根

词法

□ 词的构成

- 屈折(inflectional): 词在组句时发生的形态变化
 - foot → feet
 - book+s → books
 - I → me
 - 我 → 我们
- 派生(derivational): 构词法

词的构成

□ 构词法

- 派生：加“缀” (affixation)
 - endanger, 第一, 老师
 - excitement, modernization, 作者, 石头
 - improperly
 - 来得及, 糊里糊涂

词的构成

□ 构词法

- 复合(compounding)
 - horseman, lip-read, snapshot, highway
 - 奴隶, 火车, 提高, 顶针, 眼熟

词的构成

□ 构词法

- 词类转化(conversion)

- can **n → v**

- walk **v → n**

- poor **adj → n**

- free **adj → v**

词的构成

□ 构词法

- 拼缀法(blending)

- motor + hotel = motel

- breakfast + lunch = brunch

- situation + comedy = sitcom

- medical + care = medicare

- work + welfare = workfare

词的构成

□ 构词法

- 逆成法(backformation)
 - **edit from editor**
emote from emotion

词的构成

□ 构词法

- 缩略(abbreviation)和截短(shortening)
 - NMT, 中石油
 - ad. from adverb or advertisement

词的构成

□ 构词法

- 拟声(onomatopoeia), 重叠(reduplication), 比拟(analogy)...
 - plop, crack, 布谷
 - blah-blah, tiptop, 悄悄, 研究研究
 - 螺丝, 龙头
 - white-dollar, 白领

词的构成

□ 语法范畴(grammatical category)

- 具有相似或相同语法特性的词聚集出来的类别
- 可以是语法类别，比如词类(词性，POS)
- 可以从更细的角度去划分
 - 数 (number): book-books, 我-我们
 - 性 (gender): waiter-waitress, he-she
 - 格 (case): I-me-my
 - 时 (tense): I worked...
 - 人 (person): He plays football every Sunday

词的构成

- **体 (aspect)**——动作与时间之间的关系，尤指完成、延续或重复等状态: I have worked..., 去过, 吃了
- **态 (voice)**: 主动语态, 被动语态
- **式 (mood)**: 陈述、虚拟、祈愿、祈使
- **级 (degree of comparison)**: 比较级, 最高级
- **主 (volition)**: 动词的自主与不自主: 你看! ✓
*你看见! ✕
- **及物 (transitivity)**: 不及物、及物、双及物动词

外国人学汉语

- 我没听了
- 应该是“我没听过”（“没”和“了”不能同时用）
- 你吃饭了吗？我没吃过
- 应该是“我还没吃”（“还”表示比较近的时间内）
- 你迟到了吗今天？我还没迟到
- 。 。 。 。

- 前天我不去过大学
- 应该是“前天我没去过大学”（过去时要用“没”表示否定，和“过”搭配）
- 上学期，我没全部的考试过
- 应该是“上学期，我全部的考试都没过”（。。。 “考试没过”是个固定搭配）
- 那为什么要加“都”
- 表示全部
- 可以说“我全部的考试全部没过”吗？
- 有点奇怪

外国人学汉语

- 解释：“考试没过”的“过”是动词“通过”的意思，“上学期，我没考过试”过是助词，表示上文说过的“经历”。
- 送你个句子——
“估计你的现代汉语考试就从来没过过”
自己揣摩。

体会一下自己掌握汉语的过程和学英语的区别

词的构成

□ 语言学根据词的形态把语言分为四大类

分析语

- **孤立语**：词根构词，缺乏形态标志，语序和虚词来表现词间关系。如汉语。

综合语

- **黏着语**：词内有专门表示语法意义的附加成分，一个附加成分表达一种语法意义。如芬兰语，日语。
- **屈折语**：用词的形态变化来表示语法关系，一个形态成分可以表示若干种不同的语法意义。如英语、德语。
- **多式综合语**：动词前后附加各种表示词汇和语法意义的成分，构成相当于一个句子的动词。

词的构成

□ 语言学根据词的形态把语言分为四大类

分析语

- **孤立语**：词根构词，缺乏形态标志，语序和虚词来表现词间关系。如汉语。

综合语

- **黏着语**：词内有专门表示语法意义的附加成分，一个附加成分表达一种语法意义。如芬兰语，日语。
- **屈折语**：用词的形态变化来表示语法关系，一个形态成分可以表示若干种不同的语法意义。如英语、德语。
- **多式综合语**：动词前后附加各种表示词汇和语法意义的成分，构成相当于一个句子的动词。

Morpheme-to-word ratio

词的构成

□ 孤立语

钱是没有问题	是钱没有问题
问题是没有钱	是没钱有问题
有钱是没问題	是有钱没问題
没有钱是问题	有问题是没钱
问题是钱没有	没问题是有钱
钱没有是问题	没钱是有问题
钱有没有问题	

词的构成

□ 黏着语

食^ベ る “吃”（基本形、将来时）

食^ベ さ^セ る “吃” + 使役助动词 - 使/要求(某人)吃

食^ベ さ^セ ら^レ る “吃” + 使役助动词 + 被动助动词
- 被(其他人)要求(我)吃

食^ベ さ^セ ら^レ な^イ “吃” + 使役助动词 + 被动助动词 + 否定助动词 - 不被(其他人)要求(我)吃

食^ベ さ^セ ら^レ な^カ っ^タ “吃” + 使役助动词 + 被动助动词 + 否定助动词 + 过去助动词 - 曾不被(其他人)要求(我)吃

日语，每一个特定的功能都有一个特定的词缀

词的构成

□ 屈折语

汉语	俄语
我读书。	Я читаю книгу
你读书。	Ты читаешь книгу
他读书。	Он читает книгу
我们读书。	Мы читаем книгу
你们读书。	Вы читаете книгу
他们读书。	Они читают книгу

俄语动词形式因随主语人称而变化

词的构成

□ 多式综合语

Aliikusersuillammassuaanerartassagaluarpaalli.

aliiku-sersu-i-llammas-sua-a-nerar-ta-ssa-galuar-paal-li

“不过，他们会说他是个伟大的娱乐圈人，但……”

一段西格陵兰语，一个动词就是一句话

词法

- 词的构成
- 词法规则
- 词法分析

词法规则

□ 词法规则：词构成或转换的形式化表述

- 一个正词法(orthographic rules)的例子

$$\varepsilon \rightarrow e / \left\{ \begin{array}{c} x \\ s \\ z \end{array} \right\} \wedge \text{---} s\#$$

当词汇有一个以x(或s或z)为结尾的语素，而下一个语素是s时，插入一个e。

词法规则

□ 词法规则：词构成或转换的形式表述

- 一组形态学规则

(V ROOT ?r SUBCAT ?s VFORM pres AGR 3s) →

(V ROOT ?r SUBCAT ?s VFORM base IRREG - PRES-) + S

(V ROOT ?r SUBCAT ?s VFORM pres AGR {1s 2s 1p 2p 3p}) →

(V ROOT ?r SUBCAT ?s VFORM base IRREG - PRES-)

词法

- 词的构成
- 词法规则
- 词法分析

词法分析

□ 词法分析

- 英语等屈折语的词法分析

- **tokenization**

- **stemming** (词干还原)

- **lemmatization** (词典化)

- 汉语等孤立语的词法分析

- 重叠词 (**reduplication**)

- 离合词 (**separable verbs**)

- 汉语分词 (**Chinese word segmentation**)

词法分析

□ tokenization——相当于对屈折语分词

I'll see Mr. Green this afternoon. —————→

I/ will/ see/ Mr./ Green/ this/ afternoon/./

token这个词在深度学习范式下应用广泛，通常指输入网络的最小单元

Figures: 45.78, 90.5%, 2/3, 8/24/2010

“.”: I will see Mr. Green this afternoon.

“”: I'm, I'd, can't

“-”: stock-index, twenty-first

“\”: sound\graphics

Others: URL, computer program, formula, etc

词法分析

□ stemming——还原到词根

employers → employ + ~er + ~s

规则变化使用规则:

*ed → *	(worked → work)
*ed → *e	(believed → believe)
*ied → *y	(fied → fly)
*ing → *	(going → go)
*ing → *e	(having → have)
*es → *	(boxes → box)
*er → *	(taller → tall)
*ier → *y	(easier → easy)

不规则变化使用词典:

take – took – taken
child – children
good – better – best

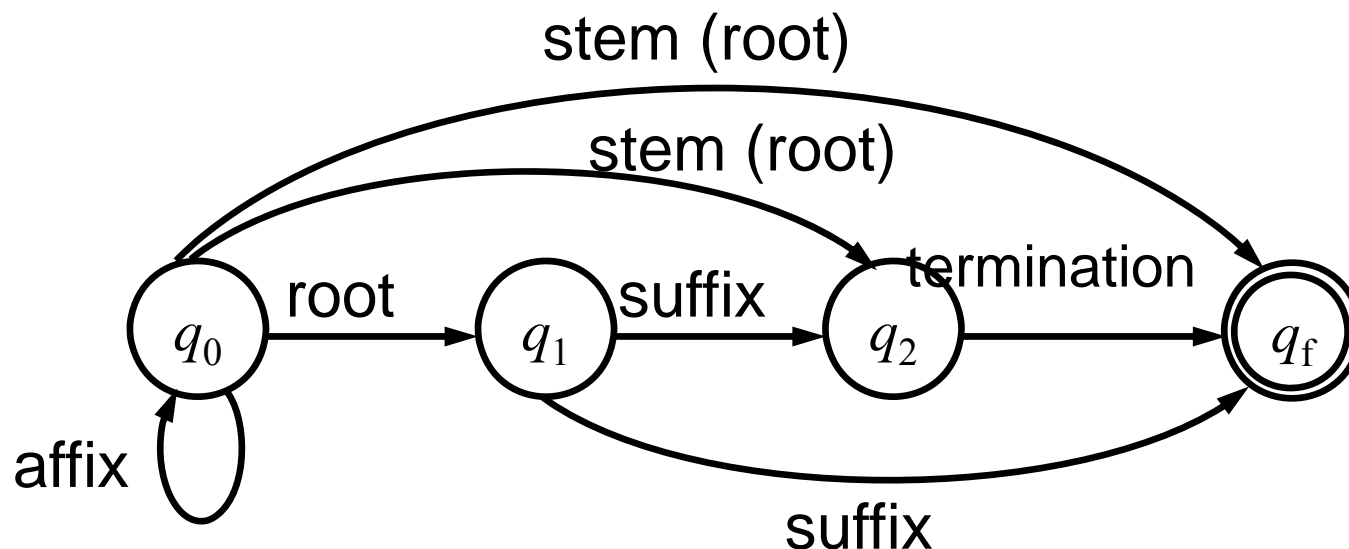
词法分析

□ lemmatization——还原到词典形式

employers \longrightarrow employer + ~s

better \longrightarrow good

屈折语形态分析主要采用规则方法以及有限状态自动机(**finite- state automaton, FSA**)



词法分析

□ 汉语的重叠词——词根或词干(或其中的一部分发生重复)

● 动词重叠

■ 单音节：V—VV, 听—听听；

■ 双音节：AB—ABAB, 形容—形容形容； AB—AABB, 来往—来来往往。

➤ 你跑跑，就知道伤没伤着了

➤ 你们多下下棋，聊聊天

➤ 再长长就该开花了

表示尝试性的，多次频繁的，持续一段时间的

词法分析

- 形容词重叠

- 单音节：J—JJ, 高一高高；

- 双音节：AB—ABAB—AABB, 高兴—高兴高兴—高兴高兴

- 名词重叠 方方面面

- 副词重叠 屡屡失败

- 数词重叠 一一取缔

- 量词重叠 场场爆满

词法分析

□ 汉语的离合词

- 看：看一看，看了看，看了一眼
- 洗澡：洗了一个澡
- 刷牙：牙刷了吗？
- 担心：担什么心

词法分析

□ 一种解决办法——借助动词的虚化：动词本身没有多少实际意义，靠后面的结伴词

➤ took a bath, make a speech, take a look, give a smile, have a shower, do the cooking

● 洗澡：洗了一个澡（DO了一个洗澡）

刷牙：牙刷了吗？（DO刷牙了吗？）

担心：担什么心（DO什么担心）

词法分析

□ 汉语重叠词、离合词的处理规则

VV---v [form: VV] << V---v

V **了** V---v [form: V **了** V] << V---v

V—V---v [form: V—V] << V---v

VVN---v [form: VVN] << VN---v

V **过** —N---v [form: V **过** —N] << VN---v

V **了** —N---v [form: V **了** —N] << VN---v

AA---a(v) [form: AA] << A---a(v)

AABB---a(v) [form: AABB] << AB---a(v)

ABAB---a(v) [form: ABAB] << AB---a(v)

词法分析

□ 汉语分词

- 分词规范——什么是一个“词”很难说清楚
- 分词歧义
- OOV

汉语分词



汉语分词

各位理工男们，如果你的女朋友问你，我的
卷发棒在哪里？

你一定要回答她，你的卷发棒就棒在特别配
你的气质。

汉语分词

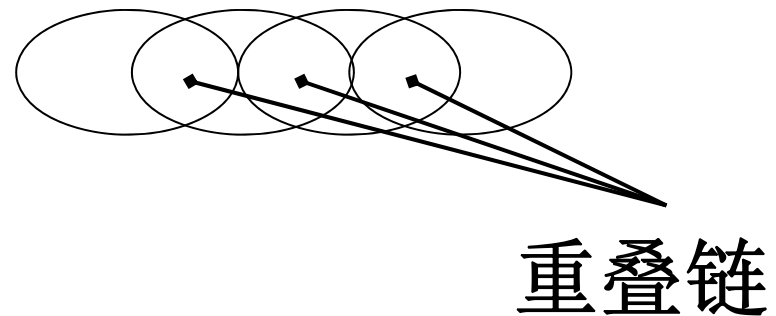
□ 分词歧义

- 重叠歧义(overlapping ambiguity)

结 合 成

投 资 产 业 结 构

以 上 海 外 资 金 融 机 构 为 核 心



汉语分词

□ 分词歧义

- 组合歧义(combination ambiguity)

你 (家) (用) 的什么样的电脑?

(家用) 电脑。

你的 (牙) (刷) 了吗? 我的 (牙刷) 不见了。

(把) (手) 举起来! 这个 (把手) 是木制的。

汉语分词

□ OOV (out of vocabulary, 未登录词)

- 未出现在系统词表中的词(unknown words, UNK)——计算机不认识的词
- 人名地名机构名
 - 伊丽莎白·亚历山德拉·玛丽·温莎，苏格兰巴尔莫勒尔堡
- 新词，术语
 - 怨种，吃席，碳中和，长短时记忆网络
- 缩写
 - 服贸会，瓜众，yyds，OPT

真实OOV，一个术语抽取场景

当O₂体积分数位于区间($\phi(O_2) < 0.4\%$)，SO₂产气量随着O₂体积分数的增加而缓慢增加；结合ArcGIS统计分析工具，采用当前比较成熟的大数据分析技术，实现了对覆冰的趋势分析、与设计冰厚对比分析及预警；

月曰`补` } 笋、小瓷套问隙U(kV)工放电压31人V(11卜)工放电压2了.skV阀终尸100夕〔屯~~卜~,一一.~目`.~一工一一山~~山一么一1主11•毛18rZ一15划夕节二价t(声5)图3图5了0高阻值并联电刚J.J曰Ut}(1、V雇小走人入并琢电阻U(kV)亨户下一一福1,`1418t吸琳5)异阀片101:18t切5)图〔宝图8U(kV)护瓷弃/U(kV)_于冒毖一,毛高准值一许状1七吐户刁片3010艘卜二`二二`三一`足一。

由以上分析可知，壁温越高，溶液所对应的饱和溶解度越低，浓度差越大，沉积率越大，污垢单位面积的沉积速度也就越快。

节点之间以对应的热阻相连，形成了定子铁心段基本单元网络(见图1，其中热网络节点之间连线表示热阻)。

图7A区控制脉冲仿真Fig.7SimulationofcontrolpulseinsectionA表3为采用逻辑法和计数法完成上述脉冲发生时所占用的CPLD资源情况。

t????t1时刻电阻降到2??，进入燃弧阶段。

NLP与语言学

- 什么是语言
- 基本概念
- 词法基础
- 句法基础
- 语义基础
- 语用基础

补充—Chomsky的形式文法理论

□ 上世纪40年代的香农Shannon

- 把离散M过程的概率模型应用于描述语言的自动机
- 把通过媒介传输语言的行为比喻成噪声信道和解码，提出了熵，并用概率技术测定了英语字母的不等概率零阶熵
- 体现的思想是：当传递的信息不是均匀分布的时候，可以花费更小的代价，用更窄的带宽来传输它

Chomsky的形式文法理论

▣ 基于MM的随机文本生成(letters)

- Zeroth-order 马尔可夫近似(symbols equiprobable)

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
FFJEYVKCQSGHYD QPAAMKBZAACIB ZLHJQD

- First-order 一阶马尔可夫近似(symbols emitted with their frequency in English text)

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH
EEI ALHENHTTPA OOBTTVA NAH BRL

Chomsky的形式文法理论

▣ 基于MM的随机文本生成(letters)

- Second-order 马尔可夫近似 (symbol bigram probabilities)

ON IE ANTSOUTINYS ARE T INCTORE ST BE S
EAMYACHINDILONASIVETU-COOWE AT
TEASONARE FUSO TIZIN ANDYTOBE SEACE CTISBE

- Third-order 马尔可夫近似(symbol trigram probabilities)

IN NO IST LAT WHEY CRATICT FROURE BIRS
GROCID PONDENOME OF DEMONSTURES OF THE
REPTAGIN IS REGOACTIONA OF CRE

Chomsky的形式文法理论

▣ 基于MM的随机文本生成(words)

- First-order 马尔可夫近似 (word probabilities)

REPRESENTING AND SPEEDILY IS AN GOOD APT OR
COME CAN DIFFERENT NAT- URAL HERE HE THE A
IN CAME THE TO OF TO EXPERT GRAY COME...

- Second-order 马尔可夫近似(word bigram probabilities)

THE HEAD AND IN FRONTAL ATTACK ON AN
ENGLISH WRITER THAT THE CHARACTER OF THIS
POINT IS THEREFORE ANOTHER METHOD FOR...

Chomsky的形式文法理论

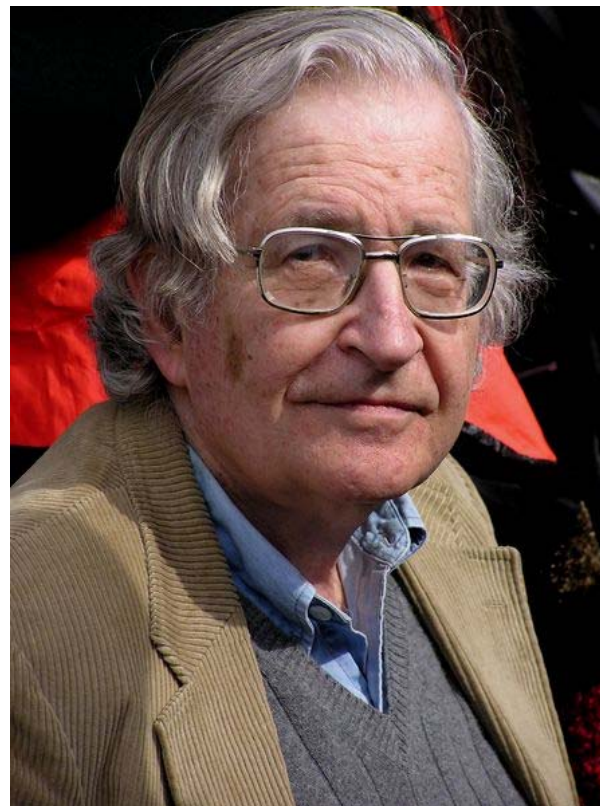
□ 乔姆斯基(Chomsky , 1928~)认为这样的模型不足以描述自然语言

□ 1957年发表

Syntactic Structures

《句法结构》

无限的语言，应该“演绎”出来



Chomsky的形式文法理论

□ 形式语法(former grammar)

- 数目有限规则的集合，这些规则可以生成语言中的合格句子，排除语言中的不合格句子

$$G = (V_n, V_t, S, P)$$

□ 形式语言(former language)

- 满足形式语法生成规则的符号串的集合

□ 自然语言和人工语言放在一个统一的平面

Chomsky的形式文法理论

➤ 举例

$$G = (\{A, S\}, \{0, 1\}, S, P)$$

$$P : S \rightarrow 0A1 \quad A1 \rightarrow A11 \quad A \rightarrow 0$$

——可以生成**001, 0011, 00111...**

➤ 生成自然语言

$$G = (V_n, V_t, S, P)$$

$$V_n = \{S, IP, VP, NP, VV, NN\}$$

$$V_t = \{\text{希望, 采取, 大使, 政府, 行动}\}$$

$$S = S$$

$$P: 1. S \rightarrow IP$$

$$2. IP \rightarrow NP \ VP$$

$$3. VP \rightarrow VV \ IP$$

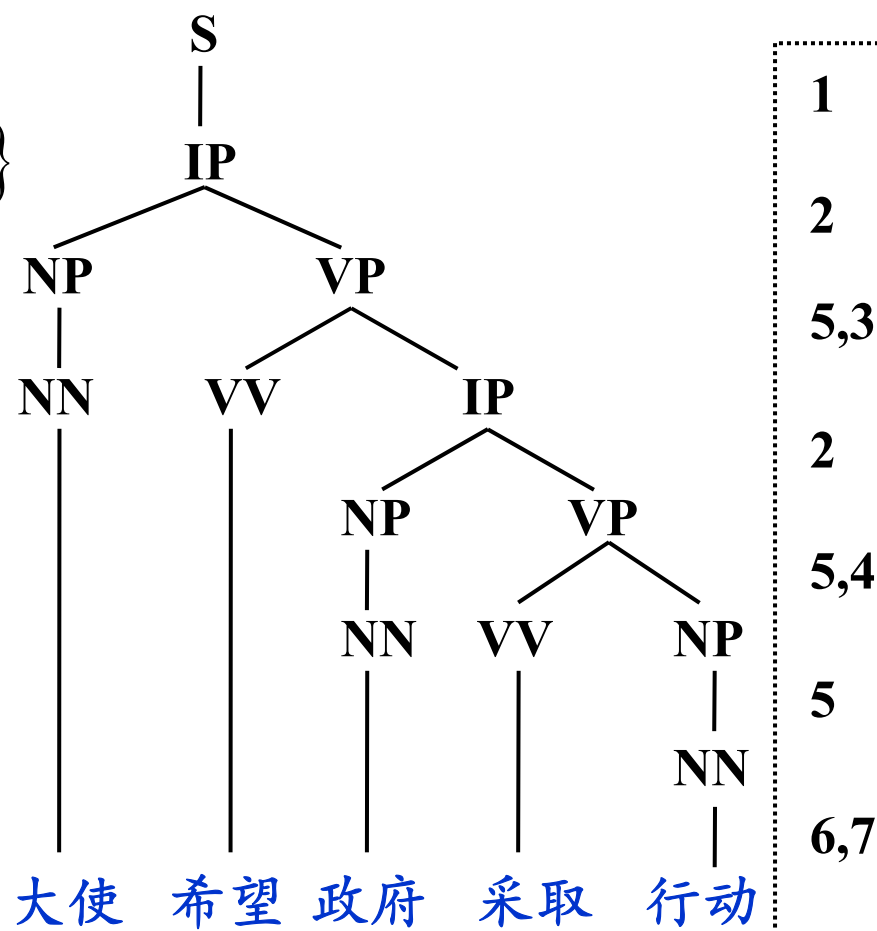
$$4. VP \rightarrow VV \ NP$$

$$5. NP \rightarrow NN$$

$$6. VV \rightarrow \text{希望} | \text{采取}$$

$$7. NN \rightarrow \text{大使} | \text{政府} | \text{行动}$$

Rules used



Chomsky的形式文法理论

- 来自王蒙小说《相见时难》中杜艳和她丈夫陈金才的一段对话，背景是陈金才拿回家的工资少了十元钱：

哪儿去了？

什么哪儿去了？

你说什么哪儿去了？

我哪儿知道你说什么哪儿去了？

你怎么会不知道我说什么哪儿去了？

你怎么知道我一定知道你说什么哪儿去了？

Chomsky的形式文法理论

➤ 来自美剧《Friends》514

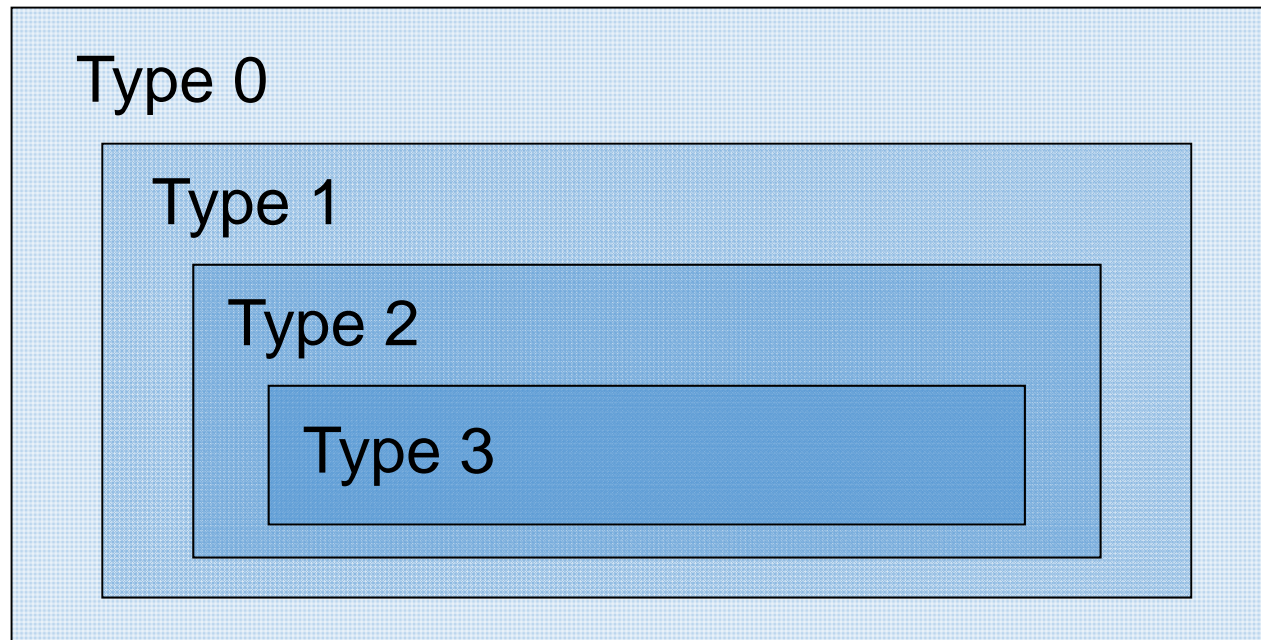
P&R: They don't know that we know.

C&M: They don't know that we know they know.

P&R: They don't know that we know they know we know.

Chomsky的形式文法理论

- 根据重写规则的不同，乔将形式语言分成具有包含关系的四个层次



Chomsky hierarchy

Type	Rule-types	Automaton	Exemplar(NL)
0: RE	$\alpha \rightarrow \beta$	通用图灵机	ATN
1: CS	$\alpha A \beta \rightarrow \alpha \gamma \beta$	线性界限自动机	TAG
2: CF	$A \rightarrow \alpha$	下推自动机(push-down)	PSG
3: FS	$A \rightarrow \begin{cases} Bx \\ x \end{cases}$	有限状态自动机(FSA)	FSTN

$$\alpha, \beta, \gamma \in (V_n \cup V_t)^*, A, B \in V_n, x \in V_t. \quad V^* = V^0 \cup V^1 \cup V^2 \dots$$

RE: 递归可枚举语言 Recursively enumerable languages

CS: 上下文相关语言 Context sensitive languages

CF: 上下文无关语言 Context-free languages

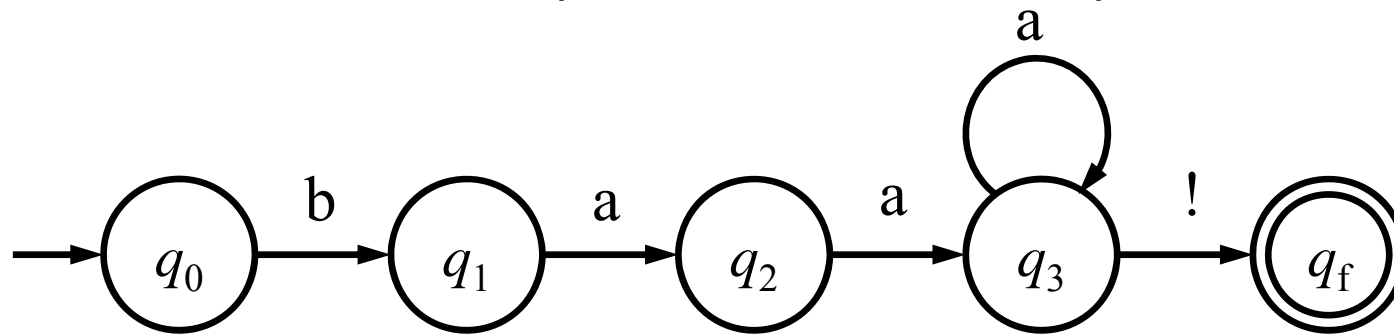
Chomsky的形式文法理论

- 那么问题是：哪一种语言可以用来描述自然语言？

Chomsky的形式文法理论

□ 有限状态文法(FSG)的有限性

- 表达FSG的FSA(有限状态自动机)



该FSA可以建模“sheeptalk”：

baa!

baaa!

baaaa!

Chomsky的形式文法理论

- FSG的生成能力有限，无法生成：

$$L_1 = \{a^n b^n\} \quad L_2 = \{a a^r\} \quad L_3 = \{a a\}$$

- 出现递归和嵌套时很难描述，而这又是自然语言一个很大的特点

Chomsky的形式文法理论

- FSG可以建模右向外围嵌套(right periphery embedding, tail recursion)
 - a. I saw [Harry swim].
 - b. I saw [John see [Harry swim]].
 - c. I saw [Anna help [John see [Harry swim]]].

Chomsky的形式文法理论

- 无法建模如下中心嵌套

a. daß ich [Heinrich schwimmen] sah.

that I Heinrich swim saw

"that I saw Henry swim"

b. daß ich [Johannes [Heinrich schwimmen] sehen] sah.

that I Johannes Heinrich swim see saw

"that I saw John see Henry swim"

c. daß ich[Anna[Johannes[Heinrich schwimmen]sehen]helfen]sah.

that I Anna Johannes Heinrich swim see help saw

"that I saw Anna help John see Henry swim"

德语的**center embedding**

Chomsky的形式文法理论

- ✓ Chomsky特别地证明了一个上下文无关语言L能够被有限状态自动机生成，当且仅当存在一个生成语言L的、没有任何中心-自嵌入(center-embedded)递归的上下文无关文法。
- 所以，描述自然语言的至少应该是上下文无关文法(CFG)

Chomsky的形式文法理论

□ CFG生成 $L_1 = \{a^n b^n\}$

$$G = (\{S\}, \{a, b\}, S, P)$$

$$P : S \rightarrow aSb \quad S \rightarrow ab$$

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow \cdots \Rightarrow a^{n-1}Sb^{n-1} \Rightarrow a^n b^n$$

□ CFG生成 $L_2 = \{a^r a^r\}$

$$G = (\{S\}, \{a, b\}, S, P)$$

$$P : S \rightarrow aa \quad S \rightarrow bb \quad S \rightarrow aSa \quad S \rightarrow bSb$$

$$S \Rightarrow bSb \Rightarrow baSab \Rightarrow babSbab \Rightarrow babbabbab$$

Chomsky的形式文法理论

□ 但 $L_3 = \{a^n a^n\}$ 依然不能生成

a. dat ik₁ Henk₂ zag₁ zwemmen₂.

that I Henk saw swim

"that I saw Henry swim"

b. dat ik₁ Jan₂ Henk₃ zag₁ zien₂ zwemmen₃.

that I Jan Henk saw see swim

"that I saw John see Henry swim"

c. dat ik₁ Anna₂ Jan₃ Henk₄ zag₁ helpen₂ zien₃ zwemmen₄

that I Anna Jan Henk saw help see swim

"that I saw Anna help John see Henry swim"

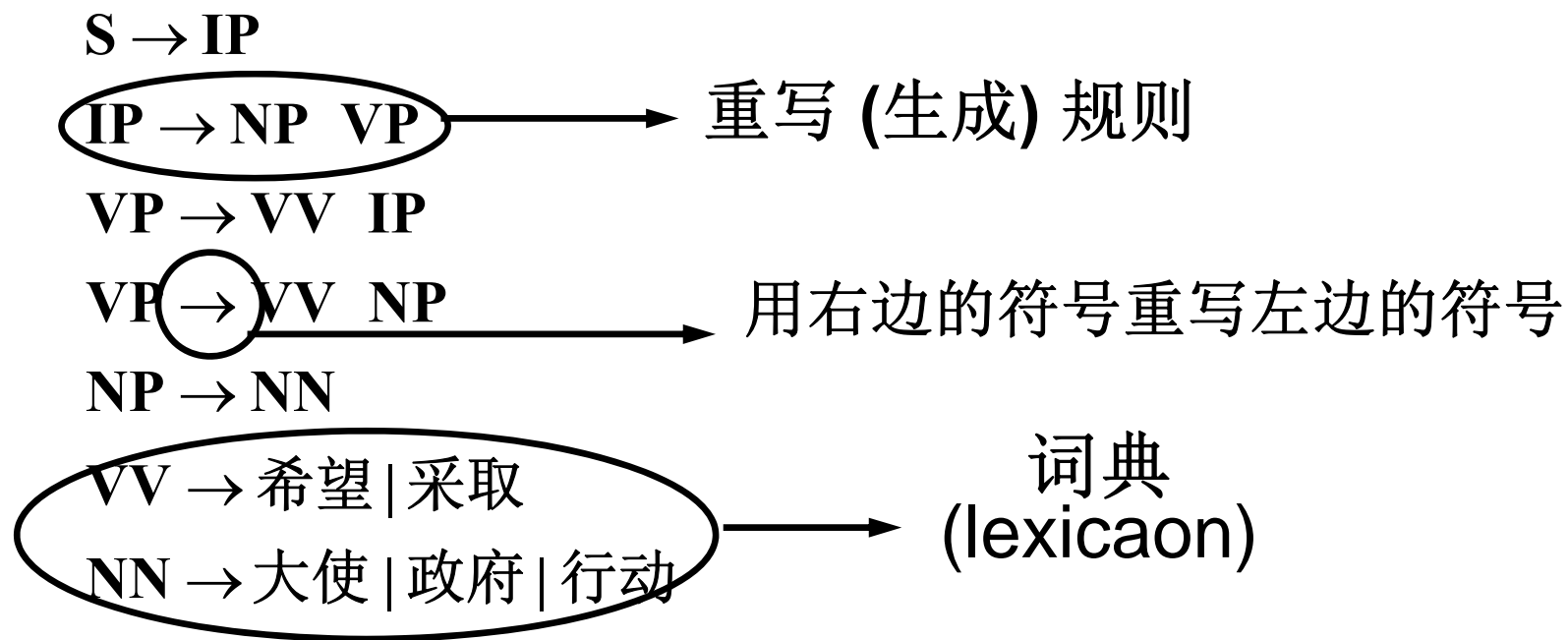
一个荷兰语的例子

Chomsky的形式文法理论

- 结论：描述自然语言的文法应该高于CFG，比如CSG
- CSG是np-complete，CFG是多项式的
- CFG是最常见的描述自然语言语法的文法

Chomsky的形式文法理论

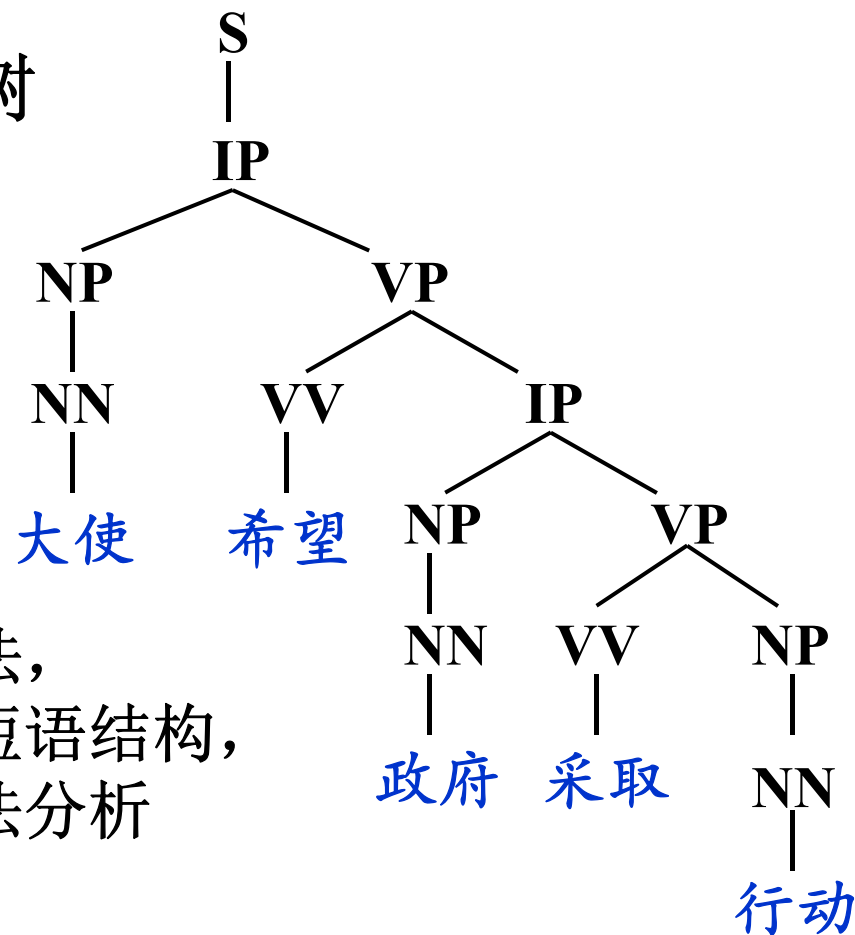
□ 基于CFG的短语结构文法



如果自然语言恰好是**CFG**描述的，那么设计合适的**CFG**，
可以生成所有的自然语言句子。

Chomsky的形式文法理论

得到句法结构树

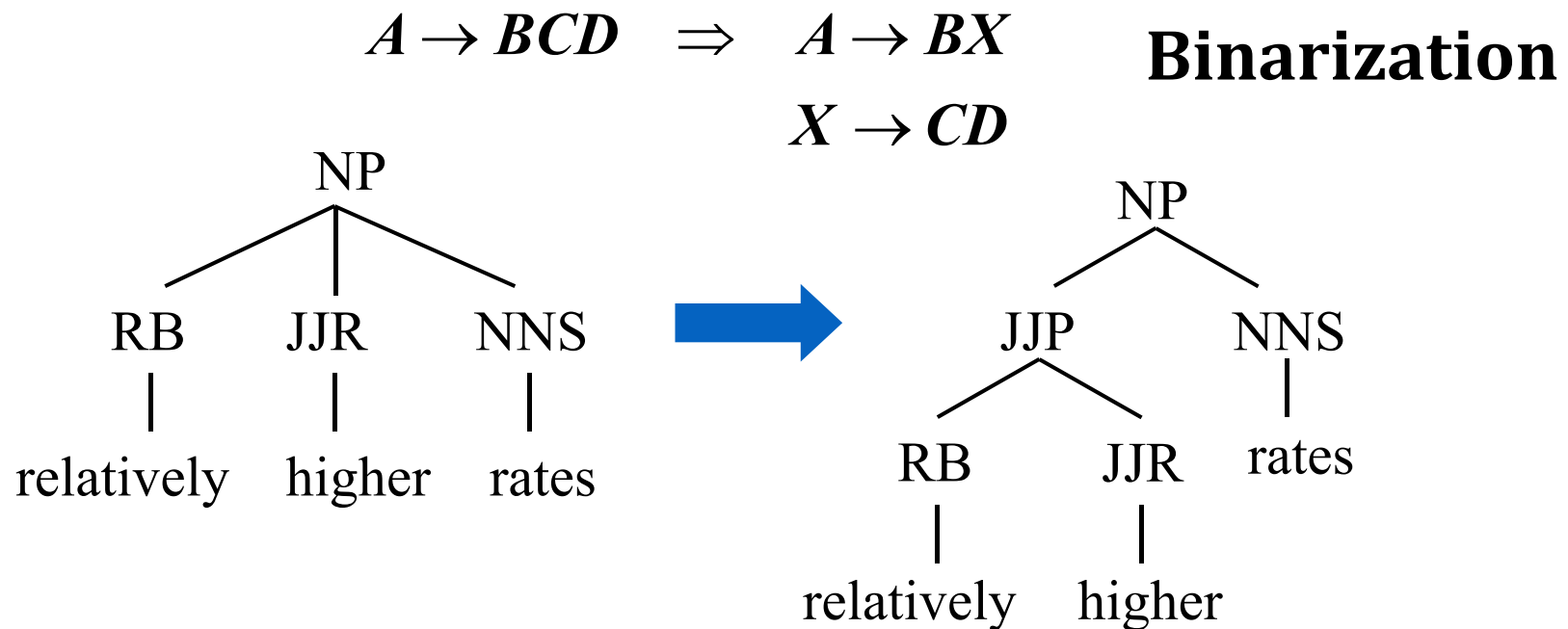


依据短语结构文法，
得到一个句子的短语结构，
称为短语结构句法分析

Chomsky的形式文法理论

□ Chomsky范式(normal form)

- 建立在句法弱等价基础上：采用二叉树来表示自然语言的句子结构。这是很多有效 parsing 算法的前提



句法

- 词汇范畴
- 句法结构
- 句法规则

词汇范畴

- 词汇范畴(lexical category): 依据词的聚合关系和组合关系划分出来的词的语法类别(词性/词类, part of speech, POS)
 - 名词、动词、形容词等

词汇范畴



词汇范畴

□ 划分依据？

- 看形态：-tion, -ness, -ment; -tive, -ful, -able。
但汉语的语素不体现句法功能
- 看词义：聪明，智慧？(意义相近但词性不同)
- 看分布，词汇在句子中的语法功能
 - 比如：名词定义为在动词前的主语或动词和介词后的宾语，但不能充当副词性的修饰词或补语。

词性标注本质上是一个句法问题

词汇范畴

□ 词性也存在一些争议

- 这本书的封面 这本书的出版
- The not observing this rule is that which the world has blamed in our satorist.
世界给予我们饥荒以作为我们无视这条规则的惩罚。
- Tom ' s winning the election was a big upset.
汤姆赢得这次选举是一个大逆转。

句法

- 词汇范畴
- 句法结构
- 句法规则

句法结构

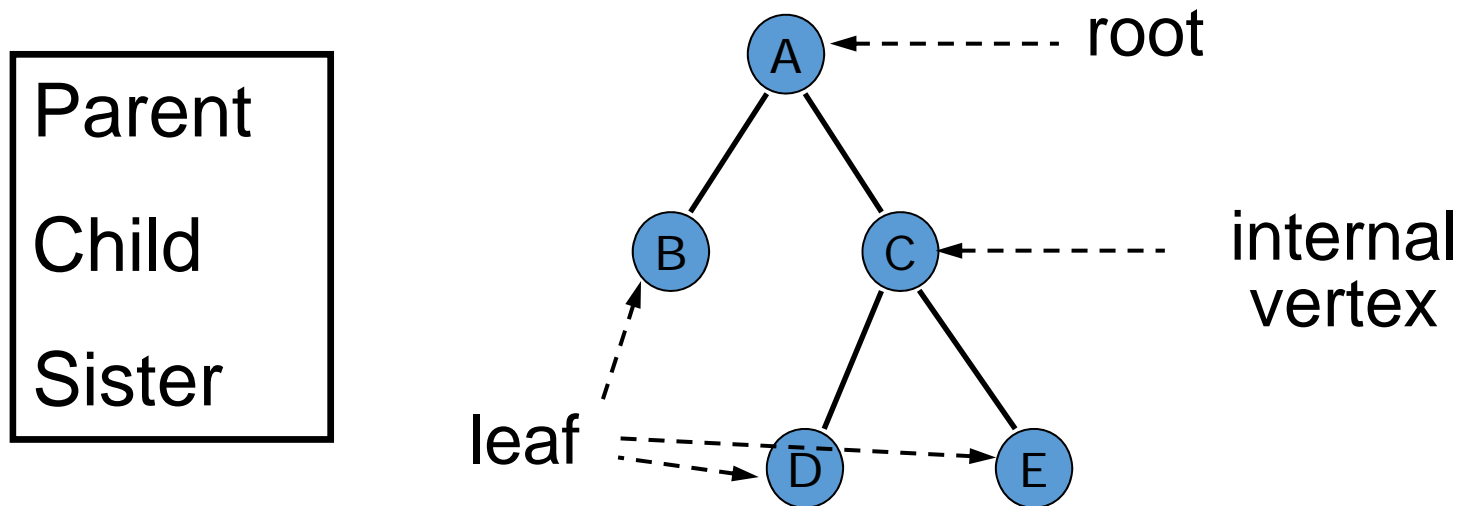
□ 句法结构(syntax)：描述法单位(句法成分)之间相互联系、相互作用的方式

“长句子” vs. “句子长”

- 结构上，都由成分“长” and “句子”构成，但关系不一样
- 功能上，“长句子”可以作为“构造”的宾语，“句子长”可以独立成句

句法结构

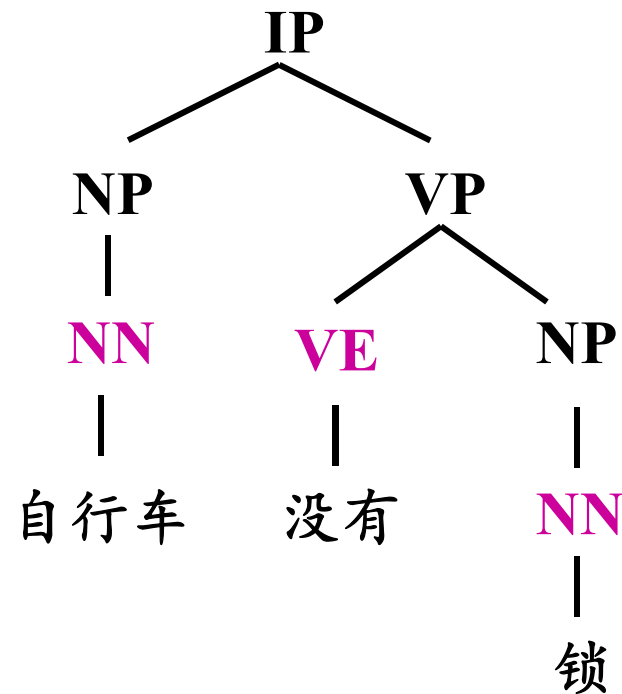
- 句子可以有线性结构(一个词挨着一个词)和层次(hierarchical)结构(嵌套)
- 句子的层次结构可以用“树”来表示
 - tree: 联通、无环、单一父节点、无向图



句法结构

□ 句子的层次结构举例

((IP (NP (NN 自行车))
 (VP (VE 没有)
 (NP (NN 锁))))))



句法

- 词汇范畴
- 句法结构
- 句法规则

句法规则

- 句法规则(syntactic rules)：一部语法里用来表示语言中符号组成和排列成句子的规律
 - 在生成语法理论中，句法规则也是生成规则(或重写规则)；递归地使用这些规则可以生成该语言的所有句子

句法规则

- 以下规则可以生成右边的树结构

IP → **NP** **VP**

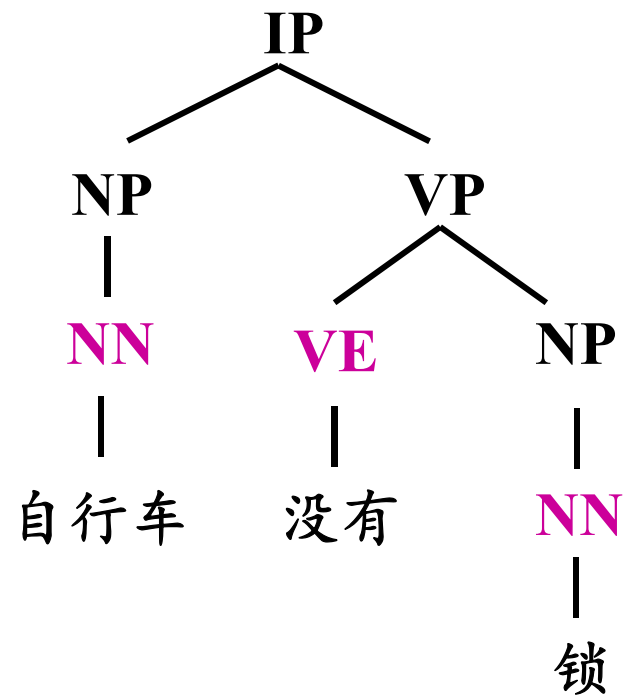
NP → **NN**

VP → **VE** **NP**

NN → 自行车

NN → 锁

VE → 没有



NLP与语言学

- 什么是语言
- 基本概念
- 词法基础
- 句法基础
- 语义基础
- 语用基础

语义

□ 语义(semantics)：简单讲，就是语言的意义(meaning)

- 语言学里讲语义，指的是句子以内的意义——句子以外的意义是语用问题

➤ 吃饭 吃食堂 吃大碗

老鼠药，感冒药，儿童药

他去修车了。

先生，您知道您的登机口吗？

Pour the water into the oil of vitriol.

后三个是
语用问题

语义

□ 广义上讲语义

- 关系的 (relational)
 - **sense** relation: synonymy (同义), antonymy (反义), hyponymy (下位)
- 组合的 (combinational)
 - 词的语义组合为句子语义(**semantic** parsing)
- 分布的 (distributional)
 - 词的语义由它的上下文决定

语义

- “语义”可能是语言学、自然语言处理中最复杂的一个概念



网页 图片 知道 视频 资讯 贴吧 文库 音乐 地图 更多»

百度为您找到相关结果约5,550,000个

搜索工具

[璠_百度汉语](#)



读音: [fán] 
部首: 王 五笔: GTOL
释义: 一种美玉。

语义检索

<https://hanyu.baidu.com> ▼

[王字旁边一个番 读什么?_百度知道](#)

2个回答 - 回答时间: 2017年10月15日

[专业] 答案:璠[fán]美玉 [beautiful jade] 璠,珣璠,鲁之宝玉也。从玉,番声。——《说文》 如圭如璠。——陆云《答顾秀才》 又如:璠玛(两种美玉);璠京(...

<https://zhidao.baidu.com/quest...> ▼

语义分析

□ semantic analysis(parsing)一般是指将自然语言句子parsing成逻辑形式(logical form)

- 相关概念:

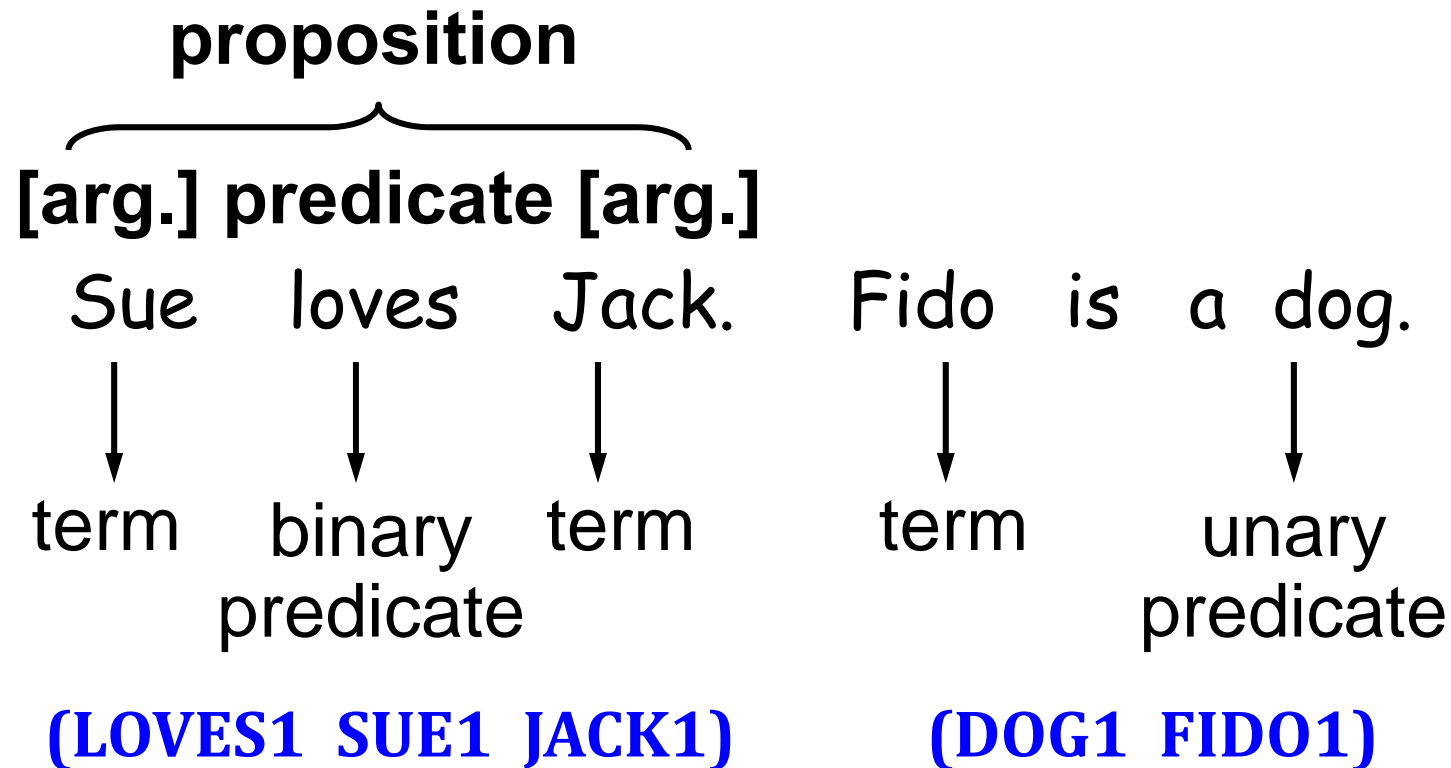
- 命题逻辑(propositional logic)

- 谓词逻辑(predicate logic)

- 谓词演算(predicate calculus)

语义分析

□ 逻辑形式语言(logical form language)



NLP与语言学

- 什么是语言
- 基本概念
- 词法基础
- 句法基础
- 语义基础
- 语用基础

语用

□ 在语言学中，对上下文(context)相关的语言的研究称为语用学(pragmatics)。

	局部上下文	全局上下文
情景 上下文	Mary是说话者，John是听者，现在是下午三点钟。他们在学生会里。他们每人有一杯咖啡，放在位于两个人之间的桌子上...	他们在美国。夏天要比春天和秋天都热。有一个月亮围绕地球在运转。
篇章 上下文	最近一个句子的分析树是 (S (NP my coffee) (VP is cold)); 下一句中的"it"指代"my coffee"	John和Mary一直在讨论他们的物理课上的一个项目。然后他们开始讨论学生会的食物质量。

语用

□ 在语言学中，对上下文(context)相关的语言的研究称为语用学(pragmatics)。

	局部上下文	全局上下文
情景 上下文	Marv是说话者，John是听者， 社交线索 ， 对话系统 。他们每人有一个咖啡杯，放在位于两个人之间的桌子上...	他们在美国。夏天要比春 世界知识的应用—知识图谱
篇章 上下文	最近一个句子的分析树是 (S (篇章问题 e) (VP is cold)) ; 下一句中的"it"指代"my coffee"	John和Marv一直在讨论他们 Topic问题— 一个项目。 主题模型等 论学生会成员们那里。

篇章

- 篇章(discourse)：句子，命题，言语行为，话轮等构成的连贯的序列
- 和长度无关，只要表达了一个特定场景下的特定语义

Fire! Help!

Just give the right order: ① ④ ⑧ ⑦ ③ ⑥ ⑤ ②

① Chinese scientists have successively **discovered** ten archaeopteryx-type fossils in the Beipiao region of Liaoxi this year.

② There are not only completely preserved **skulls** and **wings** but also waist **belts**, back **limbs** and several **feathers**.

③ **However**, the study of archaeopteryx has **never stopped**.

④ **This find** has attracted attention from home and abroad, and has been called "one of the most important finds since archaeopteryx fossils were found in Germany in **1862**".

⑤ **They** are a rich source of **material**.

⑥ **Today**, the archaeopteryx-type fossils China discovered are close to and similar in age, form and structure to Germany's archaeopteryx -- it was approximately 142.5 million years or more ago and belonged to the Mesozoic Oxfordian.

⑦ **Over the past 100 years**, the world has discovered altogether 8 archaeopteryx findings from the same period, and **only** confined to this area.

⑧ **In 1862** at the earliest finding of archaeopteryx in the state of Bavaria, Germany, it stirred the whole world.

对应的汉语篇章

中国科学家今年先后在辽西北票地区发现十枚始祖鸟类化石。这一发现引起国内外关注，被称作是“继一八六二年德国始祖鸟化石发现以来最重要的发现之一”。

一八六二年德国巴伐利亚州最早发现始祖鸟时，曾轰动全球。一百多年来，世界同时代的始祖鸟总共发现8块，且仅局限于这一地区。但是，关于始祖鸟的研究，却一直源源不断。今天，中国发现的始祖鸟类化石，其时代和形态构造都与德国始祖鸟相近和相似——它距今约一亿四千二百五十多万年，属中生代晚侏罗世；其材料丰富，不仅有保存完整的头骨和翅膀，还有腰带和后肢以及多枚羽毛。

篇章

- 国防部新增一位发言人。言论自由是一种基本人权。几年后，他的生活又基本上恢复了原来的样子。蜡梅原来不是梅花。
- A: How did you like the performance?
B: It was a nice theatre.

篇章

□ Give “me” what ?

- Tell me John's grade in *CSC271*.
Give me it in *MTH444* as well.
Give me Mike's in *MTH444* too.

翻译

冬天来了，春天还会远吗？

—If winter comes, can spring be far behind?

孔子的家里很穷，但是他从小就认真读书，刻苦学习。二十多岁的时候，做了个小官。他很有学问，办事认真，工作出色，三十岁左右，就已经很出名了。

—Confucius's family was very poor, but he studied hard even in early childhood. **He** became a petty official in **his** early twenties. By the age of thirty he had already earned a high reputation **for** his profound knowledge, his hard work, and his outstanding performance.

MT 1

- 1 Chine scrambled research on 16 key technical
- 2 These techniques are from within headline everyones boosting science and technology and achieving goals and contend of delivered on time bound through achieving breakthroughs in essential technology and complimentarity resources . national

BLEU: 0.224 (1-gram:7, 2-gram:0, 3-gram:2, 4-gram:1)

LC: 0.107 (number of lexical cohesion devices: 5) 斜体的部分

Human assessment: 2.67

MT 2

- 1 China is accelerating research 16 main technologies
- 2 These technologies are within the important realm to promote sciences and technology and achieve national goals and must be completed in a timely manner through achieving main discoveries in technology and integration of resources .

BLEU: 0.213 (1-gram:5, 2-gram:3, 3-gram:2, 4-gram:1)

LC: 0.231 (number of lexical cohesion devices: 9)

Human assessment: 4.33

Reference

- 1 China Accelerates Research on 16 Main Technologies
- 2 These technologies represent a significant part in the development of science and technology and the achievement of national goals. They must be accomplished within a fixed period of time by realizing breakthroughs in essential technologies and integration of resources.

篇章

□ 篇章的两大问题

- Cohesion 衔接
- Coherence 连贯

衔接

□ Cohesion is a way of getting text to
“hang together as a whole” .

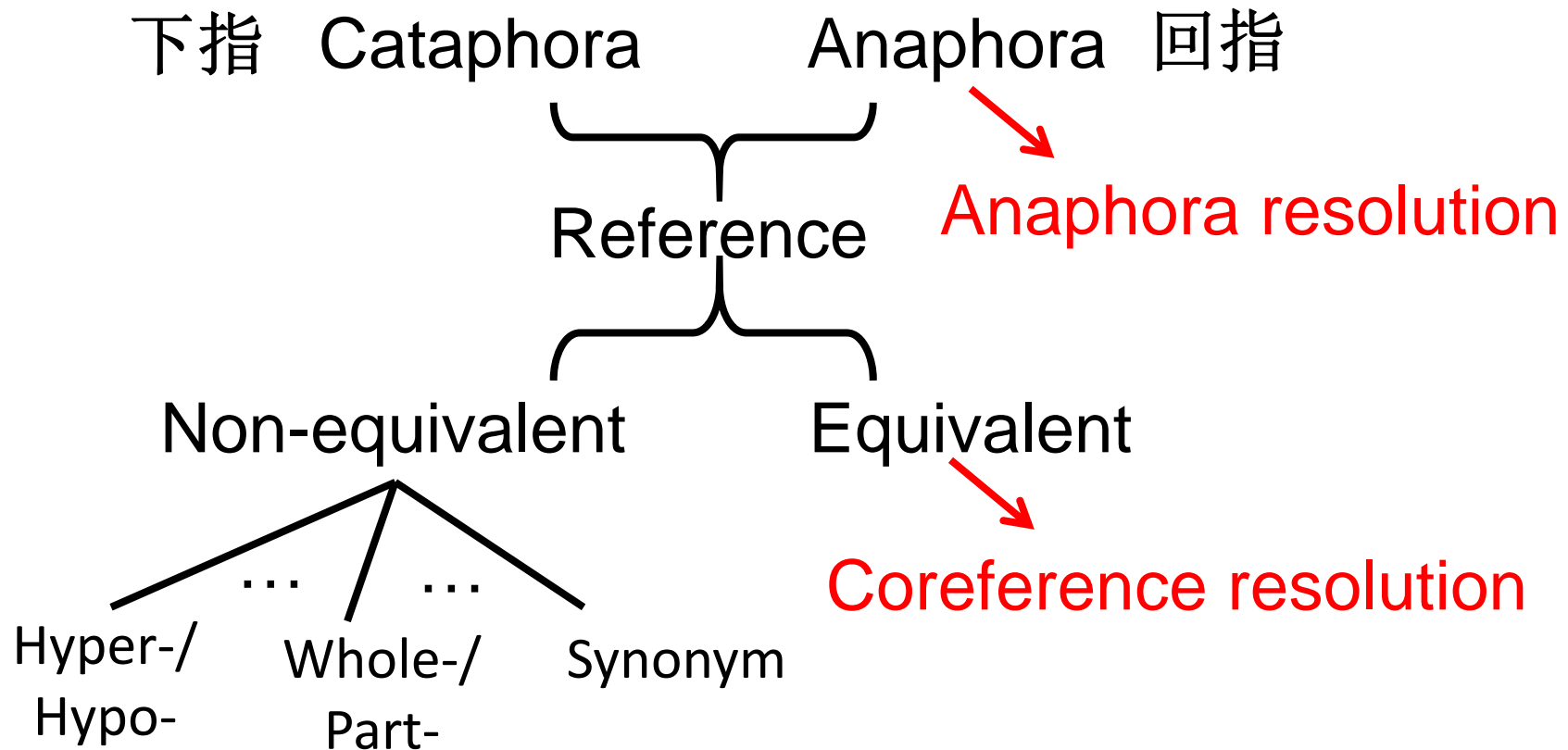
——Halliday and Hasan

□ 五种手段

- Reference (指代)
- Substitution (替换)
- Ellipsis (省略)
- Conjunction (连接)
- Lexical cohesion (词汇衔接)

衔接

□ 指代和指代消解



衔接

- 小张很聪明。他总是能很快算出答案。
- 想讨好他父亲的人，争先为小张开门。
- 花丛中跑出一只小花猫。这只猫是老王家的。
- Microsoft Corp. announced its new CEO yesterday. Microsoft side...
- 沈阳矿山机器集团公司的领导大胆创新，...，有效遏制住了经济滑坡，公司产值以平均每年33%...
- 门口停着好些三轮车，许多车夫在那里闲站着。

——《半生缘》

衔接

□ 替换

- Above is an example. Let's give you another *one*.

衔接

□ 省略

- *pro*忽然收住，赵伯韬摇摇身体站起来，
*pro*从烟匣中取一枝雪茄衔在嘴里，*pro*
又将那烟匣向立玉亭面前一推，*pro*做了个
“请吧”的手势，*pro*便又埋身在沙发里，
*pro*架起了腿，*pro*慢慢地擦火柴，*pro*
燃那枝雪茄。——矛盾《子夜》
- A: Do you understand?
B: Do you?

衔接

□ 连接

- a. 普京本星期初曾经表示他将邀请南国总统米洛舍维奇和南国反对党总统候选人科什图尼察到莫斯科举行谈判。
- b. 但是科什图尼察表示由于南国国内的政治危机，他无法离开南斯拉夫。
- c. 而米洛舍维奇对普京的提议则是没有反应。

衔接

□ 词汇衔接

- In front of me lay a **virgin** crescent cut out of **pine bush**. A dozen houses were going up, in various stages of construction, surrounded by hummocks of dry earth and stands of precariously tall **trees** nude halfway up their **trunks**. They were the kind of trees you might see in the mountains.

词汇链: {virgin, pine, bush, trees, trunks, trees}

□ 连接和词汇衔接

（语文高考试题）给下面语句排序：

- ①因为较弱的电磁辐射，也会对人的神经系统与心血管系统产生一定的干扰。
- ②人的大脑和神经会产生微弱的电磁波，当周围电器发出比它强数百万倍的电磁波时，人的神经活动就会受到严重干扰。
- ③即使在不太强的电磁波环境中工作和生活，人也会受到影响。
- ④如果长时间出于这种强电磁波的环境中，人会出现头痛，注意力不集中、嗜睡等症状，强电磁辐射会使心血管疾病加重、神经系统功能失调。

A. ④①②③

B. ②③①④

C. ④③②①

D. ②④③①

连贯

□ 连贯 : coherence is the reason why a discourse is a discourse...

- 内部连贯(interiorly): 衔接, 显式的, 隐式的
- 外部连贯(exteriorly): 情景, 背景知识

➤ 国防部新增一位发言人。言论自由是一种基本人权。几年后, 他的生活又基本上恢复了原来的样子。蜡梅原来不是梅花。

➤ A: How did you like the performance?
B: It was a nice theatre.

篇章

□ 记住

- 衔接的五大要素：指代、替换、省略、连接、词汇衔接
- 连贯；内部连贯，外部连贯

□ 通常，遇到篇章相关问题

- 如果给定篇章，则假定是衔接和连贯的，可以：指代消解，省略恢复，discourse parsing
- 又或者，需要判断是否衔接或连贯
- 或要求生成连贯的文本

回顾NLP的主要困难

□ 如何消歧？

知识
Knowledge

□ 我们需要两个方面的知识

- 语言知识 (linguistic knowledge)
- 世界知识 (world knowledge)

NLP的三大类方法(范式)都在考虑如何获取和利用知识

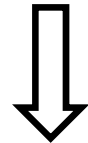
大类	细分类	特点	举例
算法 (动态的)	浅层学习	线性模型	SVM、CRF
	深度学习	非线性模型	RNN、CNN
	NLP算法	跟语言知识密切相关	CYK
知识 (狭义，显性)	语言知识	词典、规则库	WordNet、HowNet、大词林
	常识知识	很难从文本中挖到	CYC
	世界知识	可以从文本中挖到	知识图谱
数据 (广义，隐性)	有标注	专家标注、众包	Penn TreeBank
	无标注	原始语料	《人民日报》
	伪数据	最大	情感分析、 WSD 、篇章结构分析

知识的来源——刘挺(哈工大)

知识

□ 知识(狭义)表示方法：

- KRL
- 逻辑学：一阶逻辑，描述逻辑
- 心理学和语言学：语义网络



知识图谱

