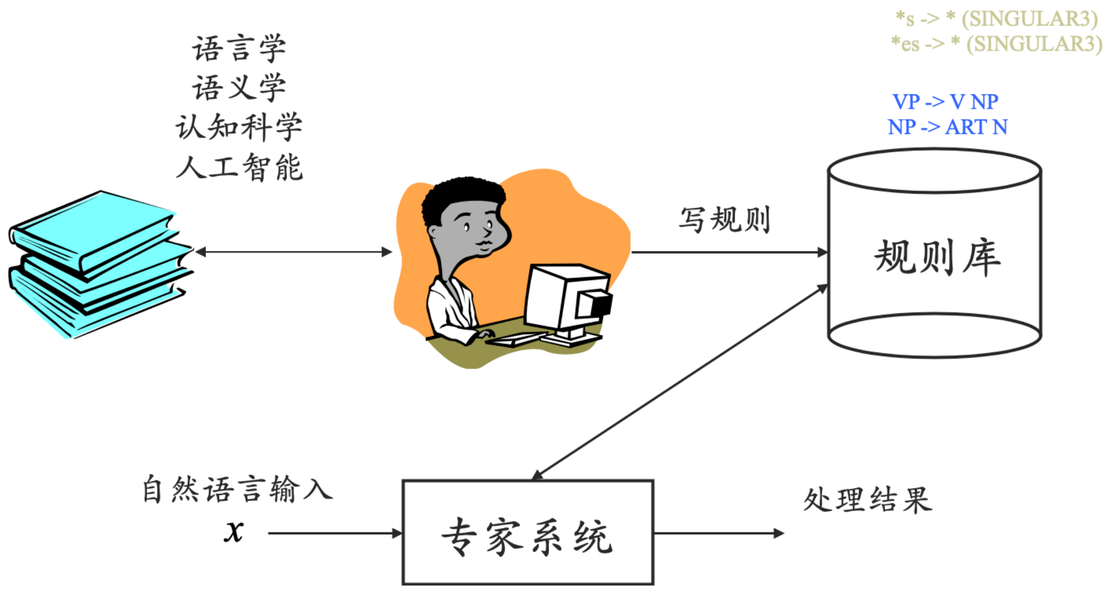
基于规则的自然语言处理方法

以词法分析为例

戴新宇，南京大学

规则方法

- 强调对语言知识的理性整理
- 受计算语言学理论指导
- 基于规则的知识表示和推导
- 语言规则（数据）与程序分离，程序体现为规则语言的解释器！



自然语言分类

基于形态结构

- 分析型语言
  - 词形变化很少
  - 没有表示词的语法功能的附加成分，由词序和虚词表示词之间的语法关系
  - 汉语、藏语等
- 黏着型语言
  - 有词形变化
  - 词的语法意义（功能）由附加成分表达
  - 芬兰语、日语等
- 屈折型语言
  - 有词形变化
  - 词的语法意义由词的形态变化来表示
  - 英语、德语、法语等
- 另外，还可以按SVO型（主－动－宾）、VSO型（动－主－宾）和SOV 型(主－宾－动) 分类

以日本语"食べる"为例：

食べ る Tabe ru - "吃" (基本形、将来时)

食べ させ る Tabe sase ru - "吃" + 使役助动词 - 使/要求(某人)吃

食べ させ られ る Tabe sase rare ru - "吃" + 使役助动词 + 被动助动词 - 被(其他人)要求(我)吃

食べ させ られ ない Tabe sase rare nai - "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

食べ させ られ な かった Tabe sase rare na katta - "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

词法分析

- 形态还原（针对英语、德语、法语等）
  - 把句子中的词还原成基本词形，作为词的其它信息（词典、个性规则）的索引
- 分词（针对汉语、日语等）
  - 识别出句子中的词
- 命名实体识别
  - 人名
  - 地名
  - 机构名
- 词性标注
  - 为句子中的词标上预定义类别集合（标注集）中的类。

词（word）

以英语为例

- 词是语言中最小的能独立运用的单位，也是语言信息处理的基本单位。
- 构词特点
  - 屈折变化：词尾和词形变化，词性不变。如：
    - study, studied,studied,studying
    - speak,spoke,spoken,speaking
  - 派生变化：加前缀和后缀，词性发生变化。如：
    - friend,friendly,friendship,...
  - 复合变化：多个单词以某种方式组合成一个词。
    - blackboard, playboy,homemade, air-conditioner.....

.

形态还原

以英语为例

- 还原规则
  - 通用规则：变化有规律
  - 个性规则：变化无规律

- 英语“规则动词”还原
  - *s -> * (SINGULAR3)
  - *es -> * (SINGULAR3)
  - *ies -> *y (SINGULAR3)
  - *ing -> * (VING)
  - *ing -> *e (VING)
  - *ying -> *ie (VING)
  - *??ing -> *? (VING)
  - *ed -> * (PAST)(VEN)
  - *ed -> *e (PAST)(VEN)
  - *ied -> *y (PAST)(VEN)
  - *??ed -> *? (PAST)(VEN)

形态还原

以英语为例

- 还原规则
  - 通用规则：变化有规律
  - 个性规则：变化无规律

- 英语不规则动词还原
  - went -> go (PAST)
  - gone -> go (VEN)
  - sat -> sit (PAST) (VEN)

汉语分词

Tokenization

- 分词是指根据某个分词规范，把一个“字”串分成“词”串。
- 分词规范
  - 难以确定何谓汉语的“词”
    - 单字词与语素的界定：猪肉、牛肉
    - 词与短语（词组）的界定：黑板、黑布
  - 信息处理用现代汉语分词规范：GB-13715（1992）
  - 具体系统可根据各自的需求制定规范

汉语分词

切分歧义

- 交集型歧义字段
  - ABC切分成AB/C或A/BC
  - 如：“和平等”
    - “独立/自主/和/平等/独立/的/原则”
    - “讨论/战争/与/和平/等/问题”
- 组合型歧义字段
  - AB切分成AB或A/B
  - 如：“马上”
    - “他/骑/在/马/上”
    - “马上/过来”
- 混合型歧义
  - 由交集型歧义和组合型歧义嵌套与交叉而成
  - 如：“太平淡”（组合型、交集型）
    - “这/墙/抹/得/太/平/了”（组合型）
    - “即使/太平/时期/也/不/应该/放松/警惕”（组合型）
    - “这/篇/文章/写/得/太/平淡/了”（交集型）

- 伪歧义与真歧义
  - 伪歧义字段指在任何情况下只有一种切分
    - “为人民”只有一种切分：“为/人民”，如：“为/人民/服务”
    - 根据歧义字段本身就能消歧
  - 真歧义字段指在不同的情况下有多种切分
    - “从小学”可以有多种切分：
      - “从小/学”，如：“从小/学/电脑”（“从小”是切分成“从小”还是“从/小”要根据分词规范！）
      - “从/小学”，如：“他/从/小学/毕业/后”
    - 根据歧义字段的上下文来消歧

分词方法

规则方法

一般通过分词词典和分词规则库进行分词。主要方法有：

- 正向最大匹配(FMM)或逆向最大匹配(RMM)
  - 从左至右(FMM)或从右至左(RMM)，取最长的词
  - 会忽略“词中有词”的现象：“幼儿园 地 节目”
- 双向最大匹配
  - 分别采用FMM和RMM进行分词
  - 如果结果一致，则认为成功；否则，
  - 采用消歧规则进行消歧（交集型歧义）：
- 正向最大、逆向最小匹配
  - 发现组合型歧义
- 逐词遍历匹配
  - 在全句中取最长的词，去掉之，对剩下字符串重复该过程
- 设立切分标记
  - 收集词首字和词尾字，把句子分成较小单位，再用某些方法切分
- 全切分
  - 获得所有可能的切分，选择最大可能的切分

分词方法

规则方法　　统计方法？？？

- 利用歧义字串、前驱字串和后继字串的句法、语义和语用信息：
  - 句法信息
    - “阵风”：根据前面是否有数词来消歧。“一/阵/风/吹/过/来”、“今天/有/阵风”
  - 语义信息
    - “了解”：“他/学会/了/解/数学/难题”（“难题”一般是“解”而不是“了解”）
  - 语用信息
    - “拍卖”：“乒乓球拍卖完了”，要根据场景（上下文）来确定
- 规则的粒度
  - 基于词（个性规则）
  - 基于词类、词义（共性规则）

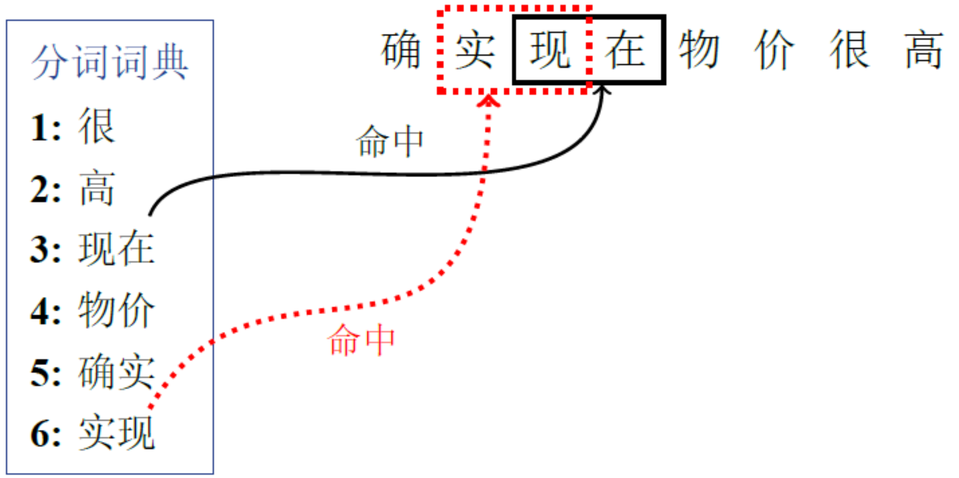


图 2.9: 交叉型分词歧义

[极限测试]

汉语分词

- 专有词 (实体识别)
  - 时间、地点、机构
- 领域词汇：过氧化氢酶
- 数字
- Email
- 新词
- ……

```
^\w+([-+.]\w+)*@\w+([-.]\w+)*\.\w+([-.]\w+)*$
```

正则表达式

Regular Expression, RE

- 一种可以高效、简洁地描述处理特定符号串（词法单元）的模式表示方法
- 字母表Σ上的正则表达式的定义
  - 基本部分
    - ε 是一个正则表达式，L(ε)={ε}
    - 如果a是Σ上的一个符号，那么a是正则表达式，L(a)={a}
  - 归纳步骤：
    - 选择：(r) | (s)，L((r) | (s))=L(r) U L(s)；
    - 连接：(r)(s)，L((r)(s))=L(r)L(s) ；
    - 闭包：(r)*，L((r)*)=(L(r))*；
    - 括号：(r)，L((r))=L(r)
  - 运算的优先级：* > 连接符 > |
  - (a)|((b)*(c))可以改写为 a|b*c

正则表达式示例

Regular Expression, RE

- $\Sigma=\{a,b\}$
- $L(a|b) = \{a,b\}$
- $L((a|b)(a|b)) = \{aa,ab,ba,bb\}$
- $L(a^*) = \{\varepsilon,a,aa,aaa,aaaa,.....\}$
- $L((a|b)^*) = \{\varepsilon,a,b,aa,ab,ba,bb,\ aaa,aab,.....\}$
- $L(a|a^*b) = \{a,b,ab,aab,aaab,...\}$

正则表达式的性质

- 等价性
- 如果两个正则表达式r和s表示同样的语言，则r=s
- 代数定律

定律	描述
$r\|s = s\|r$	|是可以交换的
$r\|(s\|t) = (r\|s)\|t$	|是可结合的
$r(st) = (rs)t$	连接是可结合的
$r(s\|t) = rs\|rt;\ (s\|t)r = sr\|tr$	连接对|是可分配的
$\epsilon r = r\epsilon = r$	ϵ是连接的单位元
$r^* = (r\|\epsilon)^*$	闭包中一定包含ϵ
$r^{**} = r^*$	* 具有幂等性

正则定义

- 对正则表达式命名，使表示简洁。

$$d_1 \rightarrow r_1$$
$$d_2 \rightarrow r_2$$
$$\cdots$$
$$d_n \rightarrow r_n$$

- 各个d_i不在字母表Σ中，且名字都不同
- 每个r_i都是$\Sigma \cup \{d_1, d_2, \ldots, d_{i-1}\}$上的正则表达式

正则定义示例

C语言的标识符集合

letter_ ➔ *A* | *B* | … | *Z* | *a* | *b* | … | *z* | _

digit ➔ 0 | 1 | … | 9

id ➔ letter_(letter_|digit)*

正则表达式的扩展定义

- 基本运算符：并 连接 闭包
- 扩展运算符
- 一个或多个：r+ , 等价于rr*
- 零个或一个： r?， 等价于ε |r
- 字符类 [abc]等价于a|b|c, [a-z]等价于a|b|...|z

$$
\begin{aligned}
\textit{letter\_} &\rightarrow [\texttt{A-Za-z\_}] \\
\textit{digit} &\rightarrow [\texttt{0-9}] \\
\textit{id} &\rightarrow \textit{letter\_}\ (\ \textit{letter\_} \mid \textit{digit}\ )^*
\end{aligned}
$$

正则表达式的扩展定义

表达式	匹配	例子
c	单个非运算符字符 c	a
\c	字符 c 的字面值	*
“s”	串 s 的字面值	“**”
.	除换行以外的任何字符	a.*b
^	一行的开始	^abc
$	行的结尾	abc$
[s]	字符串 s 中的任何一个字符	[abc]
[^s]	不在串 s 中的任何一个字符	[^abc]
r*	由和 r 匹配的零个或多个串连接成的串	a*
r^+	由和 r 匹配的一个或多个串连接成的串	a+
r?	零个或一个 r	a?
r{m,n}	最少 m 个，最多 n 个 r 的连接	a{1,5}
r_1r_2	r_1后加上r_2	ab
$r_1 \mid r_2$	r_1或r_2	a|b
(r)	与 r 相同	(a|b)
r_1/r_2	后面跟有r_2时的r_1	abc/123

词性标注

- 为句子中的词标上预定义类别集合（标注集）中的类，为后续的句法/语义分析提供必要的信息
  - 标注体系
  - 标注方法
- 词的分类
  - 按形态和句法功能（句法相关性）
  - 按表达的意思（语义相关性）
  - 兼顾上述二者
- 兼类词
  - 一个词具有两个或者两个以上的词性
  - 英文的Brown语料库中，10.4%的词是兼类词。例如：
    - The back door
    - On my back
    - Promise to back the bill
- 汉语兼类词，例如：
  - 把门锁上， 买了一把锁
  - 他研究…， 研究工作
- 汉语词的兼类更多？与所采用的分类体系是否有关？
- 为什么要分类？分类带来的问题？

1 名词	5 代词	9 量词
2 动词	6 介词	10 助词
3 形容词	7 连词	11 感叹词
4 副词	8 数词	12 拟声词

一个以义为纲的词汇分类体系

——《现代汉语分类词典》*

洪桂治　苏新春

英语词的词类

- 开放类（open class）
  - Nouns
    - 句法上：可有限定词、可作物主、有复数形式
    - 语义上：人名、地名和物名
  - Verbs
    - 句法上：几种词形变化
    - 语义上：动作、过程（一系列动作）
  - Adjectives
    - 句法上：修饰Nouns等
    - 语义上：性质
  - Adverbs
    - 句法上：修饰Verbs等
    - 语义上：方向、程度、方式、时间

- 封闭类（closed class，function words）
  - Determiners
  - Pronouns
  - Prepositions
  - Co Auxiliary verbs
  - Particles（if、not、...）
  - Numerals njunctions

词性标注方法

- 规则方法
  - 词典和规则提供候选词性
  - 消歧规则进行消歧
- 统计方法 （后续课程详细讲解）
  - 选择最可能的标注
  - 训练用语料库（已标注）
- 基于错误驱动转换学习的方法
  - 统计学习规则
  - 用规则方法进行标注