第六章 词法分析与词性标注

No.95, Zhongguancun East Road
Beijing 100080, China

http://www.ia.ac.cn
Tel. No.: +86-10-6255 4263

6.1 概述

词是自然语言中能够独立运用的最小单位，是语言信息处理的基本单位。

自动词法分析就是利用计算机对自然语言的形态 (morphology) 进行分析，判断词的结构、类别和性质。

词性或称词类（Part-of-Speech, POS）是词汇最重要的特性，是连接词汇到句法的桥梁。

6.1 概述

❑ 不同语言的词法分析

曲折语（如，英语、德语、俄语等）：用词的形态变化表示语法关系，一个形态成分可以表示若干种不同的语法意义，词根和词干与语词的附加成分结合紧密。

词法分析：词的形态变化分析，即词的形态还原。

分析语（孤立语）（如，汉语）：分词。

黏着语（如，日语）：分词＋形态还原

6.2 英语的形态分析

- 基本任务
  - 单词识别
  - 形态还原

6.2 英语的形态分析

❑ 英语单词的识别

例 (1) Mr. Green is a good English teacher.

(2) I'll see prof. Zhang home after the concert.

识别结果：

(1) Mr./ Green/ is/ a/ good/ English/ teacher/.

(2) I/ will/ see/ prof./ Zhang/ home/ after/ the/ concert/.

6.2 英语的形态分析

- 英语中常见的特殊形式的单词识别

(1) prof., Mr., Ms. Co., Oct. 等放入词典；

(2) Let's / let's => let + us

(3) I'am => I + am

(4) {it, that, this, there, what, where}'s =>

{it, that, this, there, what, where} + is

(5) can't => can + not; won't => will + not

6.2 英语的形态分析

(6) {is, was, are, were, has, have, had}n't =>
{is, was, are, were, has, have, had} + not

(7) X've => X + have;

X'll=> X + will; X're => X + are

(8) he's => he + is / has => ?

she's => she + is / has => ?

(9) X'd Y => X + would (如果 Y 为单词原型)

=> X + had (如果 Y 为过去分词)

6.2 英语的形态分析

❑ 英语单词的形态还原

1. 有规律变化单词的形态还原

1) -ed 结尾的动词过去时，去掉ed；

*ed → *　　(e.g., worked → work)

*ed → *e　　(e.g., believed → believe)

*ied → *y　　(e.g., studied → study)

6.2 英语的形态分析

2) -ing 结尾的现在分词，

*ing → * (e.g., developing → develop)

*ing → *e (e.g., saving → save)

*ying →*ie (e.g., die →dying)

3) -s 结尾的动词单数第三人称；

*s → * (e.g., works → work)

*es → * (e.g., discuss → discusses)

*ies → *y (e.g., studies → study)

6.2 英语的形态分析

4) -ly 结尾的副词

*ly → *　　(e.g., hardly → hard)

… …

5) –er/est 结尾的形容词比较级、最高级

*er → *　　(e.g., cold → colder)

*ier → *y　　(e.g., easier → easy)

……

6.2 英语的形态分析

6) s/ses/xes/ches/shes/oes/ies/ves 结尾的名词复数，ies/ves 结尾的名词还原时做相应变化：

bodies → body, shelves → shelf,

boxes → box, etc.

7) 名词所有格 X's, Xs'

6.2 英语的形态分析

2. 动词、名词、形容词、副词不规则变化单词的形态还原

一建立不规则变化词表

例：choose, chose, chosen

axis, axes

bad, worse, worst

6.2 英语的形态分析

3. 对于表示年代、时间、百分数、货币、序数词的数字形态还原

1) 1990s → 1990，标明时间名词；

2) 82th →去掉 th 后，记录该数字为序数词；

3) $200 → 去掉$，记录该数字为名词(200美圆);

4) 98.5% → 98.5%作为一个数词

6.2 英语的形态分析

4. 合成词的形态还原

1) 基数词和序数词合成的分数词， e.g., one-fourth 等。

2) 名词＋名词、形容词＋名词、动词＋名词等组成的合成名词，e.g., Human-computer, multi-engine, mixed-initiative, large-scale 等。

6.2 英语的形态分析

3) 形容词＋名词＋ed、形容词＋现在分词、副词＋现在分词、名词＋过去分词、名词＋形容词等组成的合成形容词，e.g., machine-readable, hand-coding, non-adjacent, context-free, rule-based, speaker-independent 等。

6.2 英语的形态分析

4) 名词＋动词、形容词＋动词、副词＋动词构成的合成动词，e.g., job-hunt 等

5) 其它带连字符“-”的合成词，e.g., co-operate, 7-color, bi-directional, inter-lingua, Chinese-to-English, state-of-the-art, part-of-speech, OOV-words, spin-off, top-down, quick-and-dirty, text-to-speech, semi-automatically, *i*-th 等。

6.2 英语的形态分析

❑ 形态分析的一般方法

1) 查词典，如果词典中有该词，直接确定该词的原形；

2) 根据不同情况查找相应规则对单词进行还原处理，如果还原后在词典中找到该词，则得到该词的原形；如果找不到相应变换规则或者变换后词典中仍查不到该词，则作为未登录词处理。

3) 进入未登录词处理模块。

6.3 汉语自动分词概要

- 汉语自动分词的重要性
  - 自动分词是汉语句法分析的基础
  - 词语的分析具有广泛的应用（词频统计，词典编纂，文章风格研究等）
  - 文献处理以词语为文本特征
  - “以词定字、以词定音”，用于文本校对、同音字识别、多音字辨识、简繁体转换

6.3 汉语自动分词概要

- 汉语自动分词中的主要问题
  - 汉语分词规范问题（《信息处理用限定汉语分词规范（GB13715）》）
    - 汉语中什么是词？两个不清的界限：
      1）单字词与语素（词素）
      2）词与短语

    如，花草，湖边，房顶，鸭蛋，小鸟，担水，一层，翻过？

6.3 汉语自动分词概要

- 歧义切分字段处理

1、中国人为了实现自己的梦想…(交集型歧义)

中国/ 人为/ 了/ 实现/ 自己/ 的/ 梦想

中国人/ 为了/ 实现/ 自己/ 的/ 梦想

中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想

例如：“大学生”、“研究生物”、“从小学起”、“为人民工作”、“中国产品质量”、“部分居民生活水平”等等

6.3 汉语自动分词概要

- 定义：链长

一个交集型切分歧义所拥有的交集串的集合称为交集串链，它的个数称为链长。

例如，“结合成分子”→“结合”、“合成”、“成分”、“分子”均成词，交集串的集合为｛合，成，分｝，因此，链长为3。类似地，“为人民工作”歧义字段的链长为3，“中国产品质量”字段的链长为4，“部分居民生活水平”字段的链长为6。

6.3 汉语自动分词概要

2、门把手弄坏了。 (组合型歧义)

门/ 把/ 手/ 弄/ 坏/ 了/ 。

门/ 把手/ 弄/ 坏/ 了/ 。

例如，“将来”、“现在”、“才能”、“学生会”等，都是组合型歧义字段。

6.3 汉语自动分词概要

梁南元（1987）曾经对一个含有48,092字的自然科学、社会科学样本进行了统计，结果交集型切分歧义有518个，多义组合型切分歧义有42个。据此推断，中文文本中切分歧义的出现频度约为1.2次／100字，交集型切分歧义与多义组合型切分歧义的出现比例约为12：1。

6.3 汉语自动分词概要

- 未登录词的识别

1、中国人名、地名、组织机构名、外国译名等

例如：盛中国，张建国，蔡国庆，尹集成，党政法，蔡英文，水皮，彭太发生，平川三太郎，约翰·斯特朗，詹姆斯·埃尔德等。

2、新出现的词汇、术语、个别俗语等

例如：博客，禽流感等。

6.3 汉语自动分词概要

例如：

(1) 他还兼任何应钦在福州办的东路军军官学校的政治教官。

(2) 大不列颠及北爱尔兰联合王国外交和英联邦事务大臣、议会议员杰克·斯特劳阁下在联合国安理会就伊拉克问题发言。

(3) 坐落于江苏省南京市玄武湖公园内的夏璞墩是晋代著名的文学家、科学家夏璞的衣冠冢。

6.3 汉语自动分词概要

❑ 汉语自动分词的基本原则

1、语义上无法由组合成分直接相加而得到的字串应该合并为一个分词单位。（合并原则）

如：不管三七二十一（成语），或多或少（副词片语），十三点（定量结构），六月（定名结构），谈谈（重叠结构，表示尝试），辛辛苦苦（重叠结构，加强程度），进出口（合并结构）

6.3 汉语自动分词概要

2、语类无法由组合成分直接得到的字串应该合并为一个分词单位。（合并原则）

i) 字串的语法功能不符合组合规律，如：好吃，好喝，好听，好看等

ii) 字串的内部结构不符合语法规律，如：游水等

6.3 汉语自动分词概要

❑ 汉语自动分词的辅助原则

操作性原则，富于弹性，不是绝对的。

1) 有明显分隔符标记的应该切分之（切分原则）

分隔标记指标点符号或一个词。如：

上、下课 → 上/ 下课

洗了个澡 → 洗/ 了/ 个/ 澡

6.3 汉语自动分词概要

2) 附着性语（词）素和前后词合并为一个分词单位（合并原则）

如：“吝”是一个附着语素，“不吝”、“吝于”等合并成一个词；

“员”：检查员、邮递员、技术员等；

“化”：现代化、合理化、多变化、民营化等。

6.3 汉语自动分词概要

3) 使用频率高或共现率高的字串尽量合并为一个分词单位（合并原则）

如：“进出”、“收放”（动词并列）；

“大笑”、“改称”（动词偏正）；

“关门”、“洗衣”、“卸货”（动宾）；

“春夏秋冬”、“轻重缓急”、“男女”（并列）；

“象牙”（名词偏正）；“暂不”、“毫不”、“不再”、“早已”（副词并列）等

6.3 汉语自动分词概要

4) 双音节加单音节的偏正式名词尽量合并为一个分词单位（合并原则）

如：“线、权、车、点”等所构成的偏正式名词：“国际线、分数线、贫困线”、“领导权、发言权”、“垃圾车、交通车、午餐车”、“立足点、共同点、着眼点”等。

6.3 汉语自动分词概要

5) 双音节结构的偏正式动词应尽量合并为一个分词单位（合并原则）。

本原则只适合少数偏正式动词，如：“紧追其后”、“组建完成”等，不适合动宾及主谓式复合动词。

6.3 汉语自动分词概要

6) 内部结构复杂、合并起来过于冗长的词尽量切分（切分原则）

i) 词组带接尾词：太空/ 计划/ 室、塑料/ 制品/ 业

ii) 动词带双音节结果补语：看/ 清楚、讨论/ 完毕

iii) 复杂结构：自来水/ 公司、中文/ 分词/ 规范/ 研究/ 计划

iv) 正反问句：喜欢/ 不/ 喜欢、参加/ 不/ 参加

6.3 汉语自动分词概要

v) 动宾结构、述补结构的动词带词缀时：写信/ 给、取出/ 给、穿衣/ 去

vi) 词组或句子的专名，多见于书面语，戏剧名、歌曲名等：鲸鱼/ 的/ 生/ 与/ 死、那/ 一/ 年/ 我们/ 都/ 很/ 酷

vii) 专名带普通名词：胡/ 先生、京沪/ 铁路

6.4 汉语自动分词基本算法

- 有词典切分 / 无词典切分
- 基于规则分析方法 / 基于统计方法

6.4 汉语自动分词基本算法

❑ 最大匹配法（Maximum Matching, MM）

一有词典切分，机械切分

- 正向最大匹配算法 (Forward MM, **FMM**)
- 逆向最大匹配算法 (Backward MM, **BMM**)
- 双向最大匹配算法（Bi-directional MM）

句子： $S = c_1 c_2 \cdots c_n$

假设词 $:w_i = c_1 c_2 \cdots c_m$　m 为词典中最长词的字数.

6.4 汉语自动分词基本算法

◆FMM 算法描述

0）令 i=0，当前指针 p_i 指向输入字串的初始位置，执行下面的操作：

1）计算当前指针 p_i 到字串末端的字数（即未被切分字串的长度）n，if n=1，转3)。否则，令 m=词典中最长单词的字数， if $n < m$, 令 $m = n$；

6.4 汉语自动分词基本算法

2）从当前 p_i 起取m个汉字作为词 w_i，作如下判断：

a) 如果 w_i 确实是词典中的词，则在w_i 后添加一个切分标志，转 c)；

b) 如果 w_i 不是词典中的词且 w_i 的长度大于1，将 w_i 从右端去掉一个字，转2)中的 i) 步；否则（即 w_i 的长度等于1），则在w_i 后添加一个切分标志，将w_i 作为单字词添加到词典中，执行 c)；

c) 根据 w_i 的长度修改指针 p_i 的位置，如果 p_i 指向字串末端，转3），否则， $i=i+1$，返回 1）；

3）输出切分结果，结束分词程序。

6.4 汉语自动分词基本算法

例：设词典中最长单词的字数为 7。

输入字串：他是研究生物化学的。

切分过程：他是研究生物化学的。

p↑ |

… …

他/ 是研究生物化学的。

p↑ |

FMM 切分结果：他/ 是/ 研究生/ 物化/ 学/ 的/。

BMM 切分结果：他/ 是/ 研究/ 生物/ 化学/ 的/。

6.4 汉语自动分词基本算法

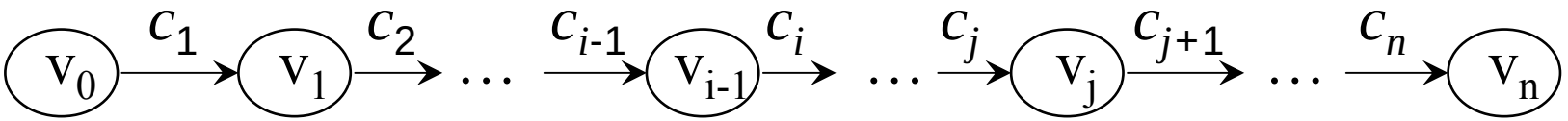
- 优点：
  - 程序简单易行，开发周期短；
  - 仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源；
- 弱点：
  - 切分歧义消解的能力差；
  - 切分正确率不高，一般在95%左右。

6.4 汉语自动分词基本算法

❑ 最少分词法（最短路径法）

◆ 基本思想

设待分字串 $S=c_1\ c_2\ldots\ldots c_n$，其中$c_i$（$i=1,2,\ldots n$）为单个的字，$n$ 为串的长度，$n\geq 1$。建立一个节点数为$n+1$的切分有向无环图G，各节点编号依次为V_0，V_1，V_2，…，V_n。



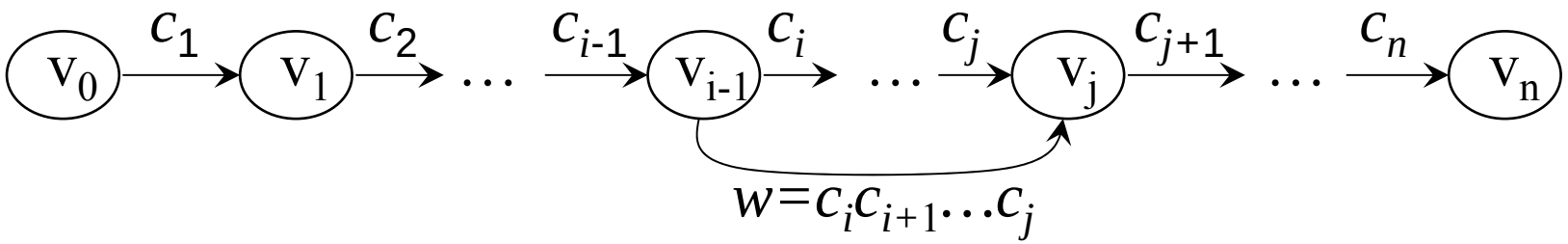
求最短路径：贪心法或简单扩展法。

6.4 汉语自动分词基本算法

◆ 算法描述：

(1) 相邻节点 v_{k-1}, v_k 之间建立有向边 $<v_{k-1}, v_k>$，边对应的词默认为 c_k (k =1, 2, …, n)。

(2) 如果 $w= c_i c_{i+1} \ldots c_j$ $(0<i<j\leq n)$ 是一个词，则节点v_{i-1}, v_j 之间建立有向边 $<v_{i-1}, v_j>$，边对应的词为 w。



(3)重复上述步骤(2)，直到没有新路径（词序列）产生。

(4)从产生的所有路径中，选择路径最短的（词数最少的）作为最终分词结果。

6.4 汉语自动分词基本算法

例：1) 输入字串：他只会诊断一般的疾病。

可能的输出：他/ 只会/ 诊断/ 一般/ 的/ 疾病。 (6)

他/ 只/ 会诊/ 断/ 一般/ 的/ 疾病。(7)

…… ……

最终结果：他/ 只会/ 诊断/ 一般/ 的/ 疾病。

2) 输入字串：他说的确实在理。

可能的输出：他/ 说/ 的/ 确实/ 在理。 (5)

他/ 说/ 的确/ 实在/ 理。 (5)

…… ……

6.4 汉语自动分词基本算法

- 优点：
  - 切分原则符合汉语自身规律
  - 需要的语言资源（词表）也不多
- 弱点：
  - 对许多歧义字段难以区分，最短路径有多条时，选择最终的输出结果缺乏应有的标准。
  - 字串长度较大和选取的最短路径数增大时，长度相同的路径数急剧增加，选择最终正确的结果困难越来越大。

6.4 汉语自动分词基本算法

❑ 基于统计语言模型的分词方法

◆ 方法描述：

设对于待切分的句子S，$W = w_1 w_2 \ldots\ldots w_k \ (1 \le k \le n)$是一种可能的切分。

$$W^* = \arg\max_W P(W \mid S)$$

$$= \arg\max_W P(W) P(S \mid W)$$

6.4 汉语自动分词基本算法

微软研究院把一个可能的词序列 W 转换成一个可能的词类序列 $C = c_1 c_2 \cdots c_N$，即：

- 专有名词的人名PN、地名LN、机构名ON分别作为一类；
- 实体名词中的日期 dat、时间tim、百分数per、货币mon等作为一类；
- 对词法派生词MW和词表词LW，每个词单独作为一类。

6.4 汉语自动分词基本算法

那么，

$$C^* = \arg\max_C P(C)P(S|C) \quad (1)$$

语言模型

生成模型

$P(C)$ 可采用3元语法：

$$P(C) = P(c_1)P(c_2|c_1)\prod_{i=3}^{N} P(c_i|c_{i-2}c_{i-1}) \quad (2)$$

6.4 汉语自动分词基本算法

生成模型在满足独立性假设的条件下，可近似为：

$$P(S \mid C) \approx \prod_{i=1}^{N} P(s_i \mid c_i) \qquad (3)$$

该公式认为，任意一个词类生成汉字串的概率只与自身有关，而与其上下文无关。例如，如果“教授”是词表词，则 $P(s_i=\text{教授}|c_i=\text{LW})=1$。

6.4 汉语自动分词基本算法

词 类	生成模型 $P(S|C)$	语言知识
词表词（LW）	若S是词表词，$P(S|LW)=1$，否则为0；	分词词表
词法派生词（MW）	若S是派生词，$P(S|MW)=1$，否则为0；	派生词词表
人名（PN）	基于字的二元模型	姓氏表，中文人名模板
地名（LN）	基于字的二元模型	地名表、地名关键词表、地名简称表
机构名（ON）	基于词类的二元模型	机关名关键词表，机构名简称表
实体名（FT）	若S可用实体名词规则集G识别，$P(S|G)=1$，否则为0。	实体名词规则集

6.4 汉语自动分词基本算法

模型的训练由以下三步组成：1)在词表和派生词表的基础上，用正向最大匹配法切分训练语料，专有名词通过一个专门模块标注，实体名词通过相应的规则和有限状态自动机标注，由此产生一个带词类别标记的初始语料；2）用带词类别标记的初始语料，采用最大似然估计方法估计统计语言模型的概率参数；3）采用得到的语言模型对训练语料重新进行切分和标注（公式(1)、(2)、(3)），得到一个刷新的训练语料。重复第2）、3）步，直到系统的性能不再有明显的提高为止。

6.4 汉语自动分词基本算法

- 优点：
  - 减少了很多手工标注知识库（语义词典、规则等）的工作
  - 在训练语料规模足够大和覆盖领域足够多的情况下，可以获得较高的切分正确率
- 弱点：
  - 训练语料的规模和覆盖领域不好把握
  - 计算量较大

6.4 汉语自动分词基本算法

❑ 基于HMM的分词方法

基本思想：

可以把输入句子S 作为HMM的输入；单词串S_w 为状态的输出，即观察序列 $S_w = w_1 w_2 \cdots w_n \quad (n \geq 1)$；词性序列 S_c 为状态序列，每个词性标记对应HMM中的一个状态 q_i，$S_c = c_1 c_2 \cdots c_n$。

6.4 汉语自动分词基本算法

利用HMM处理分词问题恰好对应于解决HMM的三个基本问题：

1）估计模型的参数；

2）对于一个给定的输入S及其可能的输出序列S_w和模型$\mu = (A, B, \pi)$，快速地计算$P(S_w \mid \mu)$。所有可能的S_w中使概率$P(S_w \mid \mu)$最大的解即为分词结果；

3）快速地选择“最优”的状态序列（词性序列），使其“最好地解释”观察序列。

6.4 汉语自动分词基本算法

模型 $\mu = (A, B, \pi)$ 中状态（词性）的数目为词性符号的个数N；从每个状态可能输出的不同符号（单词）的数目为词汇的个数M。

假设在统计意义上每个词性的概率分布只与上一个词的词性有关（即词性的二元语法），而每个单词的概率分布只与其词性相关。那么，我们就可以通过对已分词并做了词性标注的训练语料进行统计，得到如下三个矩阵：

6.4 汉语自动分词基本算法

(a) 初始状态（词性）的概率分布矩阵：

$$\pi_i = P(q_1 = c_i), \quad 1 \le i \le N$$

(b) 状态转移（词性到词性的转移）概率矩阵：

$$A = \{a_{ij}\}, \quad a_{ij} = P(q_t = c_j \mid q_{t-1} = c_i), \quad 1 \le i, j \le N$$

(c) 从状态（词性）观察到输出符号（单词）的概率分布矩阵：

$$B = \{b_j\}, \quad b_j(k) = P(S_{w_t} = w_k \mid q_t = c_j), \quad 1 \le j \le N, \quad 1 \le k \le M$$

6.4 汉语自动分词基本算法

- 基于3元统计模型的分词与词性标注一体化方法

基本思想：设句子S是由词串组成$W=w_1w_2\cdots w_n \quad (n\geq 1)$，单词$w_i$的词性标注为$t_i$，即句子$S$相应的词性标注符号序列可表达为$T=t_1t_2\cdots t_n$。那么，分词与词性标注的任务就是要在$S$所对应的各种切分和标注形式中，寻找$T$和$W$的联合概率$P(W, T)$为最优的词切分和标注组合。

6.4 汉语自动分词基本算法

$P(W, T)$ 可以由HMM 近似地表示为：

$$P(W,T) = P(W|T)P(T) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}t_{i-2}) \qquad (4)$$

生成模型

基于词性的语言模型

6.4 汉语自动分词基本算法

如果把单词序列作为HMM的中间状态，词性符号作为输出，那么，$P(W, T)$的另一种形式为：

$$P(W,T) = P(T|W)P(W) \approx \prod_{i=1}^{n} P(t_i | w_i) P(w_i | w_{i-1} w_{i-2}) \quad (5)$$

生成模型

基于词的语言模型

6.4 汉语自动分词基本算法

将上述(4)和(5)综合：

$$P^*(W,T)=\alpha\prod_{i=3}^{n}P(w_i\mid t_i)P(t_i\mid t_{i-1},t_{i-2})+\beta\prod_{i=1}^{n}P(t_i\mid w_i)P(w_i\mid w_{i-1},w_{i-2})$$

显然，这种综合模型的指导思想是希望通过调整参数α和β的值来确定两个子模型在整个分词与词性标注过程中所发挥作用的比重，从而获得分词与词性标注的整体最优。

6.4 汉语自动分词基本算法

从公式 (5) 得到的结果分析可知，$P(t_i \mid w_i)$ 对分词无帮助，且在分词确定后对词性标注又会增添偏差。因此，我们在实现这一模型时，仅取公式 (5) 中的语言模型部分，而舍弃词性标注部分，并令 α ＝1，仅保留加权系统β，于是，

$$P^{\wedge}(W,T)=\prod_{i=1}^{n}P(w_i \mid t_i)P(t_i \mid t_{i-1},t_{i-2})+\beta\prod_{i=1}^{n}P(w_i \mid w_{i-1},w_{i-2})$$

6.4 汉语自动分词基本算法

在确定β系数值时，我们根据词典中词汇w的个数和词性t的种类数目，取二者之比，即β=词典中词w的个数 / 词性t的种类数。

6.4 汉语自动分词基本算法

- 其它方法
  - 串频统计和词形匹配相结合的分词方法
  - 规则方法与统计方法相结合
  - 多重扫描法
  - 全切分方法

……

6.4 汉语自动分词基本算法

❑ 方法比较

(1) 最大匹配分词算法是一种简单的基于词表的分词方法，有着非常广泛的应用。这种方法只需要最少的语言资源（仅需要一个词表，不需要任何词法、句法、语义知识），程序实现简单，开发周期短，是一个简单实用的方法，但对歧义字段的处理能力不够强大。

6.4 汉语自动分词基本算法

(2) 全切分方法首先切分出与词表匹配的所有可能的词，然后运用统计语言模型和决策算法决定最优的切分结果。这种切分方法的优点是可以发现所有的切分歧义，但解决歧义的方法很大程度上取决于统计语言模型的精度和决策算法，需要大量的标注语料，并且分词速度也因搜索空间的增大而有所缓慢。

6.4 汉语自动分词基本算法

(3) 最短路径分词方法的切分原则是使切分出来的词数最少。这种切分原则多数情况下符合汉语的语言规律，但无法处理例外的情况，而且如果最短路径不止一条时，系统往往不能确定最优解。

(4) 统计方法具有较强的歧义区分能力，但需要大规模标注 (或预处理) 语料库的支持，需要的系统开销也较大。

6.5 未登录词识别

- 命名实体（Named Entity）（专有名词）

  人名（中国人名和外国译名）、地名、组织机构名、数字、日期、货币数量

- 其它新词

  专业术语、新词等。

6.5 未登录词识别

❑ 关于中文姓名

- 台湾出版的《中国姓氏集》收集姓氏 5544 个

  其中，单姓 3410 个，复姓 1990 个，3字姓 144 个

- 中国目前仍使用的姓氏共 737 个

  其中，单姓 729 个，复姓 8 个

- 根据我们收集的 300 万个人名统计：姓氏：974个

  其中，单姓 952个，复姓 23 个，300万人名中出现汉字4064个。 (曹文洁，2002a, 2002b)

6.5 未登录词识别

- 中文姓名识别的难点
  - 名字用字范围广，分布松散，规律不很明显。
  - 姓氏和名字都可以单独使用用于特指某一人。
  - 许多姓氏用字和名字用字（词）可以作为普通用字或词被使用，如，姓氏：于（介词），张（量词），江（名词）等；名字：建国，国庆，胜利，文革等。
  - 缺乏可利用的启发标记

6.5 未登录词识别

例如: (1) 祝贺老总百战百胜。

(2) 林徽因此时已经离开了那里。

(3) 赵微笑着走了。

(4) 南京市长江大桥。

6.5 未登录词识别

- 中文姓名识别方法
  - 姓名库匹配，以姓氏作为触发信息，寻找潜在的名字
  - 计算潜在姓名的概率估值及相应姓氏的姓名阀值，根据姓名概率评价函数和修饰规则对潜在的姓名进行筛选。

6.5 未登录词识别

设姓名 $Cname = Xm_1m_2$，其中 X 表示姓，m_1m_2 分别表示名字首字和名字尾字。

分别用下列公式计算姓氏和名字的使用频率：

$$F(X) = \frac{X\text{用作姓氏}}{\text{X出现的总次数}}$$

$$F(m_1) = \frac{m_1\text{作为名字首字出现的次 数}}{m_1\text{出现的总次数}}$$

$$F(m_2) = \frac{m_2\text{作为名字首字出现的次 数}}{m_2\text{出现的总次数}}$$

6.5 未登录词识别

字串 *Cname* 可能为姓名的概率估值：

$$P(Cname)=\begin{cases}F(X)\times F(m_1)\times F(m_2) & \text{复名情况}\\ F(X)\times F(m_2) & \text{单名情况}\end{cases}$$

姓氏 X 构成姓名的最小阀值：

$$T_{\min}(X)=\begin{cases}F(X)\times Min(F(m_1)\times F(m_2)) & \text{复名情况}\\ F(X)\times Min(F(m_2)) & \text{单名情况}\end{cases}$$

6.5 未登录词识别

姓名的评价函数：

$$f = \ln P(Cname)$$

对于特定的姓氏 X 通过训练语料得到一阈值 β_X (threshold value)，当 f 大于 β_X 时，该识别的汉字串确定为中文姓名。

6.5 未登录词识别

修饰规则：

如果姓名前是一个数字，或者与“.”字符的距离小于 2个字节，则否定此姓名。

◆ 确定潜在的姓名边界

➢ 左界规则：若潜在姓名前面是一称谓，或一标点符号，或者潜在姓名在句首，或者潜在的姓名的姓氏使用频率为100%，则姓名的左界确定。

6.5 未登录词识别

右界规则：若姓名后面是一称谓，或者是一指界动词（如，说，是，指出，认为等）或标点符号，或者潜在的姓名在句尾，或者潜在姓名的尾字使用频率为100%，则姓名的右界确定。

- 校正潜在的姓名

  依据：含重合部分的潜在姓名不可能同时成立。

  利用各种规则消除冲突的潜在姓名。

6.5 未登录词识别

❑ 中文地名识别方法

◆ 困难

- ➢ 地名数量大，缺乏明确、规范的定义。《中华人民共和国地名录》（1994）收集88026个，不包括相当一部分街道、胡同、村庄等小地方名称。
- ➢ 真实语料中地名出现情况复杂。如地名简称、地名用词与其它普通词冲突、地名是其它专用名词的一部分，地名长度不一等。

6.5 未登录词识别

- 基本资源
  - 建立地名资源知识库
    - 地名库
    - 地名用字库
    - 地名用词库
  - 建立识别规则库
    - 筛选规则
    - 确认规则
    - 否定规则

6.5 未登录词识别

- 基本方法
  - 统计模型
  - 通过训练语料选取阈值（threshold value）
  - 地名初筛选
  - 寻找可以利用的上下文信息
  - 利用规则进一步确定地名

6.5 未登录词识别

❑ 中文机构名称的识别

◆ 中文机构名称的构成

- ➢ 词法角度：偏正式（修饰格式）的复合词

  {名词|形容词|数量词|动词} + 名词

- ➢ 句法角度：“定语＋名词性中心语”型的名词短语（定名型短语）
- ➢ 中心语：机构称呼词，如：大学，学院，研究所，学会，公司等。

6.5 未登录词识别

- 中文机构名称的类型
  - 地名，如：北京大学，武汉大学
  - 人名，如：中山大学，哈佛大学
  - 学科、专业合部门系统，如：公安部，教育委员会
  - 研究、生产或经营等活动的对象，如：软件研究所，卫星制造厂
  - 上述情况的综合，如：白求恩医科大学

6.5 未登录词识别

- 大机构、团体、组织和职业的名称，如：中国人民解放军洛阳外国语学院，中国发明家学会等
- 专造的机构名，如：复旦大学，四通公司，微软研究院
- 创办、工作的方式，如：某某股份公司，中央电视大学

6.5 未登录词识别

- 机构名称识别方法
  - 找到一机构称呼词
  - 根据相应规则往前逐个检查名词作为修饰名词的合法性，直到发现非法词
  - 如果所接受的修饰词同机构称呼词构成一个合法的机构名称，则记录该机构名称
  - 统计模型

6.6 词性标注

❑ 概要

词性（part-of-speech, POS）标注（tagging）的主要任务是消除词性兼类歧义。

例如，在英语中：

1) Time **flies like** an arrow.

2) I want you to **web** our annual report.

对 Brown 语料库的统计，55%词次兼类。汉语中常用词兼类现象严重，《现代汉语八百词》兼类占 **22.5%**。

6.6 词性标注

◆ 汉语中的词性兼类现象： [赵铁军，2001]

1）形同音不同，如：“好（hao3，形容词）、好（hao4，动词）”

2）同形、同音，但意义毫不相干，如：“会（会议，名词）、会（能够、动词）”

3）具有典型意义的兼类词，如：“典型（名词或形容词）”

4）上述情况的组合，如：“行（xing2，动词/形容词；hang2，名词/量词）”

6.7 词性标注

- 标注集的确定原则：

  不同语言中，词性划分基本上已经约定俗成。

  自然语言处理中对词性标记要求相对细致。

- 一般原则：
  - 标准性：普遍使用和认可的分类标准和符号集；
  - 兼容性：与已有资源标记尽量一致，或可转换；
  - 可扩展性：扩充或修改；

6.7 词性标注

◆UPenn Treebank 的词性标注集确定原则：

- 可恢复性(recoverability)：从标注语料能恢复原词汇或借助于句法信息能区分不同的词类；
- 一致性（consistency）：功能相同的词应该属于同一类；
- 不明确性（indeterminacy）：为了避免标注者在不明确的条件下任意决定标注类型，允许标注者给出多个标记（限于一些特殊情况）。

—[Marcus et al., 1993]

6.7 词性标注

- 词性标注方法
  - 基于规则的词性标注方法
  - 基于统计模型的词性标注方法
  - 规则和统计方法相结合的词性标注方法
  - 基于有限状态变换机的词性标注方法
  - 基于神经网络的词性标注方法

6.7 词性标注

- 基于规则的词性标注方法
  - TAGGIT 系统，86种词类，3300规则

    (Bwon University)
  - 手工编写词性歧义消除规则
  - 机器自动学习规则

6.7 词性标注

- 基于规则的词性标注方法
  - 手工编写消歧规则
    - 非兼类词典
    - 兼类词典
      - 词性可能出现的概率高低排列
    - 构造兼类词识别规则

—[刘开瑛，2000]

6.7 词性标注

- 并列鉴别规则

如：体现了人民的要求（N/V ?）和愿望（N，非兼类）。

- 同境鉴别规则

如：一个优秀的企业必须具备一流的产品（名词，非兼类）、一流的管理（N/V ?）和一流的服务（N/V ?）。

6.7 词性标注

- 区别词鉴别规则（区别词只能直接修饰名词）

  如：他们搞的这次大型（鉴别词，非兼类）调查（V/N ?）历时半年。

- 唯名形容词鉴别规则（有些形容词只能直接修饰名词）

  如：重大（唯名形容词）损失（N/V ?）

  巨大（唯名形容词）影响（N/V ?）

6.7 词性标注

- 根据词语的结构建立词性标注规则
  - 词缀（前缀、后缀）规则
    - 形容词：蓝茵茵，绿油油，金灿灿，…
    - 数量词：一片片，一次次，一回回，…
    - 人名简称：李总，张工，刘老，…
    - 其它：年轻化，知识化，…{化}

      篮球赛，足球赛，…{赛}

      … …

6.7 词性标注

- 重叠词规则

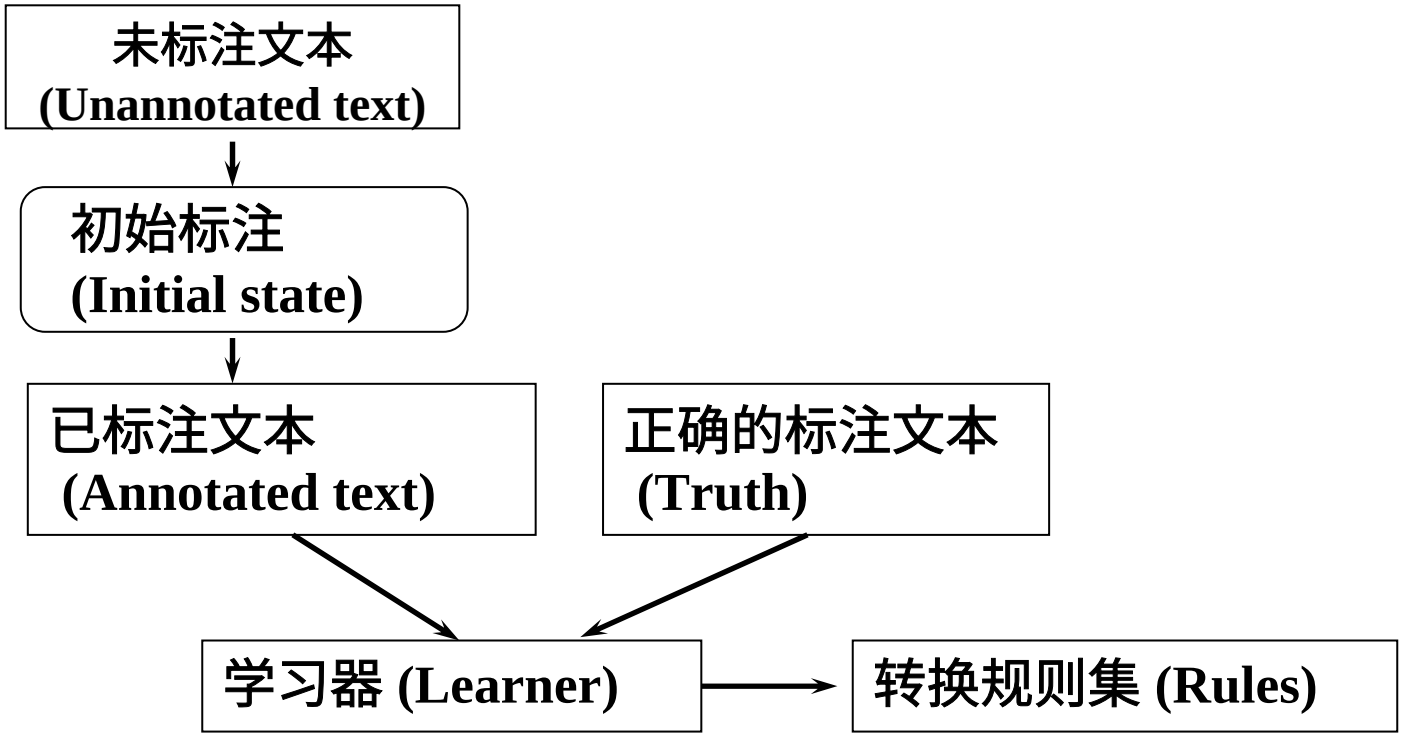
  — 看看，瞧瞧，高高兴兴，热热闹闹，…

6.7 词性标注

- 基于错误驱动的机器学习方法
  - 初始词性赋值
  - 对比正确标注的句子，自动学习结构转换规则
  - 利用转换规则调整初始赋值

—[E. Brill, 1992]

6.7 词性标注



基于转换规则的错误驱动的机器学习方法

6.7 词性标注

❑ 基于统计模型的词性标注方法

◆ 基于 *n*-gram 的语言模型

<u>应用系统</u>：1）1983年 Mashall 提出的 LOB 语料库的标注系统: CLAWS (Constituent-Likelihood Automatic Word-tagging System)

2）DeRose 对 CLAWS 改进后 VOLSUNGA 系统 (bi-gram)。

6.7 词性标注

- 基于 HMM 的词性标注方法
  - 状态集（词性序列，状态数：词类符号数）
  - 输出符号（单词序列，词汇量）
  - 初始状态概率
  - 状态转移概率
  - 符号输出概率

—[Manning, 2001] pp. 357-359:

. Jelink's Method

. Kupier's Method

6.7 词性标注

- 规则和统计相结合的词性标注方法

  规则消歧，统计概率引导

  或者统计方法赋初值，规则消歧

  —[周强，1995；张民，1998]

6.8 分词与词性标注结果评测

- 两种测试
  - 封闭测试 / 开放测试
  - 专项测试 / 总体测试

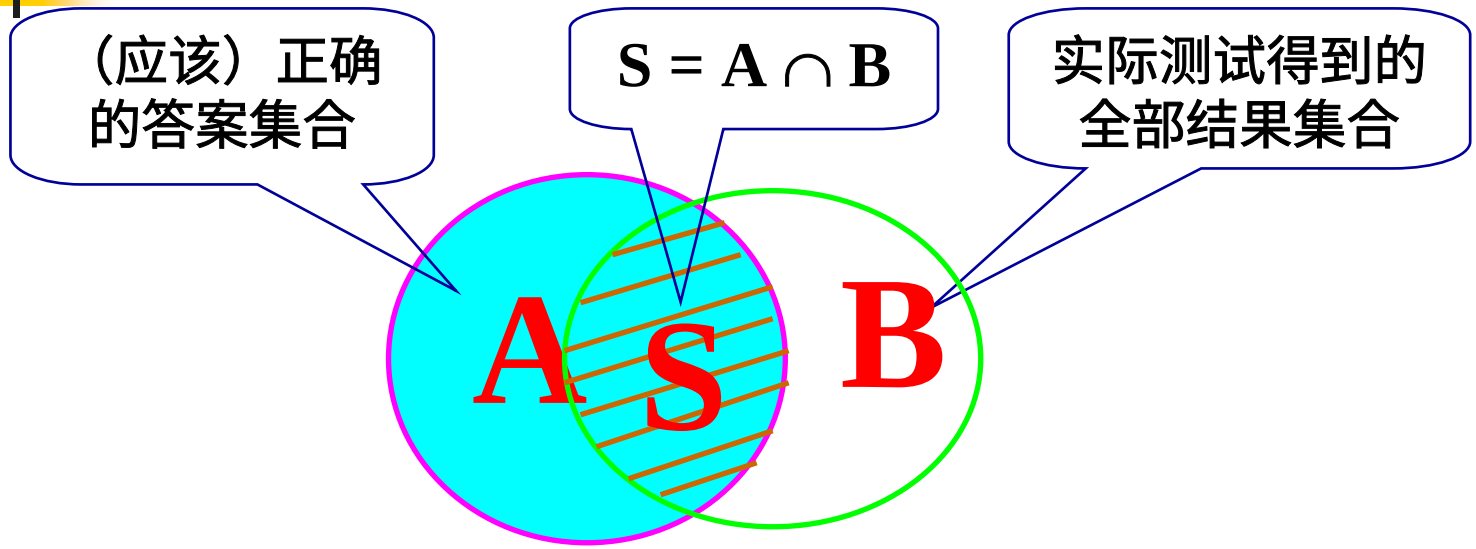
6.8 分词与词性标注结果评测

❑ 评测指标

- 正确率(Correct ratio/Precision, C): 测试结果中正确结果的个数占系统所有输出结果的比例。
- 召回率(找回率) (Recall ratio, R): 测试结果中正确结果的个数占标准答案总数的比例。
- F-测度值(F-Measure): 正确率与找回率的综合值。

$$F-measure=\frac{(\beta^2+1)\times C\times R}{\beta^2\times C+R}\times 100\%$$ 一般地，$\beta=1$。

6.8 分词与词性标注结果评测



正确率(Correct ratio): $C = \frac{S}{B} \times 100\%$

召回率(Recall ratio): $R = \frac{S}{A} \times 100\%$

本章小结

- 词法分析的任务（英语汉语有所不同）
- 英语形态分析
  - 单词识别
  - 形态还原
- 汉语自动分词
  - 汉语分词中的主要问题
  - 基本原则和辅助原则
  - 几种基本方法（MM, 最少分词法，统计法等）

本章小结

- 未登录词识别
  - 人名、地名、机构名等
- 词性标注
  - 问题（兼类、标注集、规范）
  - 方法（规则方法、统计方法、综合方法）
- 分词与词性标注结果评测
  - 正确率、找回率、F-测度值

习题

1. 设计并实现算法用于还原英语动词。
2. 设计一个有限状态自动机用于识别缩写 {he, she}'s 是 he / she has 还是 he / she is，并编写程序实现该自动机。
3. 编写程序实现汉语逆向最大分词算法（可采用有限词表），并利用该程序对一段中文文本进行分词实验，校对切分结果，计算该程序分词的正确率、召回率及F-测度。

习题

4. 设计并实现一个汉语未登录词的识别算法（可限定条件），并通过实验分析该算法的优缺点。
5. 了解目前常见的几种汉语词性标注集，比较它们的差异，并阐述你个人的观点。
6. 掌握各种词性标注方法的要点，了解目前汉语词性标注的几种主要方法。

习题

7. 试参考前人的工作，提出消除汉语自动分词中组合歧义的几点设想。
8. 阅读《信息处理用现代汉语分词规范》（中华人民共和国国家标准 GB13715），了解规范的基本内容。

Thanks

谢谢!