第二章 分 词

王峰
华东师大计算机系

主要内容

- 英语形态分析
- 汉语分词
  - 文本分词的重要性
  - 文本分词面对的问题
  - 文本分词的基本方法
  - 对文本分词质量的评价
- 小结

概 述

- 词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。

- 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，判断词的结构和类别等。

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：词的形态分析(形态还原)。
- 分析语(孤立语)（如：汉语)：分词。
- 黏着语(如：日语等)：分词＋形态还原。

英语单词的形态分析

基本任务

- 单词识别 **(tokenization)**
- 形态还原 **(lemmatization)**

英语单词的识别

例1: Mr. Green is a good English teacher.

例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/.

英语单词的识别

- 英语中常见的特殊形式的单词识别　缩写、连写

(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

英语单词的识别

(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) X'd Y => X + would (如果Y为单词原型)

=> X + had (如果Y为过去分词)

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

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

英语单词的形态还原

- 确定词的原型
  - develop, develops, developed, developing, development
- 确定句子时态、人称等

还原是为了确定词义，同时也要注意其表达的时态问题

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)**

2) -ing 结尾的现在分词

- ***ing → * (e.g., developing → develop)**
- ***ing → *e (e.g., saving → save)**
- ***ying → *ie (e.g., dying → die)**

3) -s 结尾的动词单数第三人称

- ***s → * (e.g., works → work)**
- ***es → * (e.g., discusses → discuss)**
- ***ies → *y (e.g., studies → study)**

英语单词的形态还原

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) 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,……**

7) 名词所有格 **X's, Xs'**

- **Mike's → Mike**

英语单词的形态还原

2. 动词、名词、形容词、副词不规则变化单词的形态还原
   - ✓ 建立不规则变化词表
   - 例：**choose, chose, chosen**
     axis, axes
     bad, worse, worst

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) 87th → 去掉 th 后，记录该数字为序数词；
   3) $20 → 去掉$，记录该数字为名词(20美圆)；
   4) 98.5% → 98.5 作为一个数词

英语单词的形态还原

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

英语单词的形态还原

- 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
- 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) 其他带连字符“-”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等

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。

调用nltk

- `import nltk`

  `sentence = “Tony’ s horse isn’ t from U.S.A”`

  `tokens = nltk.word_tokenize(sentence)`

  `print (tokens)`

- [‘Tony’, ”’s”, ’horse’, ‘is’, “n’t”, ‘from’, ‘U.S.A’]

还需要进一步处理

NLTK Lemmatizer

- ```
  from nltk.stem import WordNetLemmatizer
  lemmatizaer = WordNetLemmatizer()
  print(lemmatizaer.lemmatize('dogs'))
  print(lemmatizaer.lemmatize('is'))
  print(lemmatizaer.lemmatize('is',pos='v' ))
  ```
- dog
- is
- be
  ```

汉语自动分词

1 汉语自动分词的重要性

- “词”在“字”与“句”之间是隐性的单位
- 文本分词是各个层次的自然语言处理任务的基础
- 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 如词袋模型
- “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

1 汉语自动分词的重要性

简繁转换示例:

明成皇后，她是一个世纪前北韩王朝的最后一位皇后。
明成皇后，她是一個世紀前北韓王朝的最後一位皇后。

负离子陶瓷烫发机，内置负离子发射器。
負離子陶瓷燙髮機，內置負離子發射器。

每个战斗单位只有一斗米
每個戰鬥單位只有一斗米

1 汉语自动分词的重要性

- 文语转换示例
  - 1. 为达到赢球的目的，一定要注意比赛时的情绪调动与心理调节
  - 2. 他们村有三百多人种树
  - 3. 他喜欢展览馆门口播放的“小老鼠上灯台”这类欢快的儿歌
- 信息检索示例
  - 关键词：人为
  - 搜索结果：
    - 人为因素
    - 以人为本，人为什么活着

2 文本分词面对的问题

1. 什么是中文的“词”

2. 分词歧义

3. 未登录词识别

2.1 什么是“词”

- 语法学定义：能够独立运用的最小的音义结合体
  - 分词规范：结合紧密，使用稳定
  - 两个不清的界限
    - 单字词与词素，如：新华社25日讯
    - 词与短语，如：花草，湖边，房顶，鸭蛋，小 鸟，担水，一层，翻过？
- 词表定义：枚举“词型” 词典太大，计算机处理受影响
- 语料库定义：枚举“词例” 统计自然语言处理重要资源

不同的人对“词”的认识有差异

- **6**人对**100**句（**4372**字）进行人工分词，然后两两比较认同率

	M2	M3	T1	T2	T3
M1	0.77	**0.69**	0.71	**0.69**	0.70
M2		0.72	0.73	0.71	0.70
M3			**0.89**	0.87	0.80
T1				0.88	0.82
T2					0.78

平均值
0.76

Sproat R. et al., 1996, A Stochastic Finite-state Word Segmentation Algorithm for Chinese. Computational Linguistics, Vol.22, No.3, pp377-404.

分词规范

- 刘源 等（**1994**）《信息处理用现代汉语分词规范及自动分词方法》，清华大学出版社、广西科学技术出版社，**1994**年版.
- 黄居仁、陈克健等（**1997**），《信息处理用中文分词规范设计理念及规范内容》，载《语言文字应用》**1997**年第**1**期.
- 《信息处理用汉语分词规范》**GB/T13715-92**，中国标准出版社，**1993.**
- 《资讯处理用中文分词规范》，台湾中研院，**1995.**
- 《人民日报》语料库词语切分规范，北大计算语言所，**1999.**

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。
   - 例如：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）
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。
   - 字串的语法功能不符合组合规律，如：好吃，好喝，好听，好看等
   - 字串的内部结构不符合语法规律，如：游水等

2.2 文本分词中的歧义

1. 张店区大学生不看重大城市的户口本

张店区 / 大学生 / 不 / 看 / 重大 / 城市 / 的 / 户口本　　交集型歧义

张店区 / 大学生 / 不 / 看重 / 大 / 城市 / 的 / 户口本

2. 你认为学生会听老师的吗

你 / 认为 / 学生会 / 听 / 老师 / 的 / 吗　　组合型歧义

你/认为 / 学生 / 会 / 听 / 老师 / 的 / 吗

3. 只有雷人才能吸引人

只有 / 雷人 / 才能 / 吸引 / 人　　混合型歧义

只有 / 雷 / 人才 / 能 / 吸引 / 人

只有 / 雷 / 人 / 才 / 能 / 吸引 / 人

交集型歧义的链长

- 交集型歧义字段中含有交集字段的个数，称为链长
  - 链长为**1**：和尚未　　**{尚}**
  - 链长为**2**：结合成分　　**{合，成}**
  - 链长为**3**：为人民工作　　**{人，民，工}**
  - 链长为**4**：中国产品质量
  - 链长为**5**：鞭炮声响彻夜空
  - 链长为**6**：努力学习语法规则
  - 链长为7：中国企业主要求解决
  - 链长为**8**：治理解放大道路面积水

真实文本中分词歧义情况

- 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**。

- **[1]** 刘挺、王开铸，**1998**，关于歧义字段切分的思考与实验。《中文信息学报》第**2**期，**63-64**页。
- **[2]** 刘开瑛，**2000**，《中文文本自动分词和标注》，商务印书馆，**65**页。
- 交集型歧义：组合型歧义 **= 1: 22**

真实文本中分词歧义情况

- 真歧义
  - 确实能在真实语料中发现多种切分形式
  - 比如“应用于”、“地面积”、“解除了”
- 伪歧义
  - 虽然有多种切分可能性，但在真实语料中往往取其中一种切分形式
  - 比如“挨批评”、“市政府”、“太平淡”、“ 充分发挥”、“情不自禁地”

真实文本中分词歧义情况

- **78248**个交集型歧义字段**[1]**
  - 伪歧义：**94%**
  - 真歧义：**6%**
    - 多种切分均匀分布 **12%**：**A**
    - 一种切分切分占优 **88%**：**B**

A: 将信息技术**/**应用**/**于**/**教学实践

信息技术**/**应**/**用于**/**教学中的哪个方面

B: 上级**/**解除**/**了**/**他的职务

方程的**/**解**/**除了**/**零以外还有**…**

[1] 刘开瑛，**2000**，《中文文本自动分词和标注》，商务印书馆，**66-67**页。

2.3未登录词

1. 汉族人名、地名： 雪村，老张，中关村
2. 外族人名、地名： 横路静二，突尼斯
3. 中外组织机构单位名称： 联合国教科文组织
4. 商品品牌名： 非常可乐，苹果**iPad**
5. 专业术语： 有限状态自动机，三分球秒杀
6. 新词语： 蚁族，博客，非典，恶搞，给力
7. 缩略语： 人影办，两会，北医三院
8. 汉语重叠形式、离合词等： 高高兴兴，幽了他一默

2.3 未登录词

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。

待分词文本示例

不管你相不相信，美丽岛的贫困县人口中有五分之三已经超过半年领不到救济粮，其中散居全县各区的汉族人占到绝大多数。更糟糕的是，省气象部门预测即将到来的雨季降水量之大会让近半数的人无家可归。政府为此召开了紧急动员会。没想到的 是，会上一些大型国有控股企业和东南沿海民营企业的代表高高 兴兴地表示会施以援手。正在当地集训的某大洋洲国家羽毛球队参加了此次还算说得过去的动员会。

3 文本分词的基本方法

- 从字串到词串，存在着不确定性，分词的过程也就是一个降低不确定性的过程。

- 为了降低不确定性，需要为计算机提供确定的“语言知识”，比如词典、规则、经过分词处理的语料库（可从中获取词语的各项统计数据）等知识形式。

文本分词的基本方法

例：これはいぬです

词语
…
これ
はい
いぬ
です
…

根据词典进行分词：

これ/ はい/ ぬ/です

これ/ は/ いぬ/です

文本分词的基本方法

基于词典的分词方法
- 最大匹配法
- 最短路径法
- 最大概率法
- ……

自顶向下“分”词

基于字序列标注的方法

深度学习关注的方法

自底向上“合”词

最大匹配法
(Maximum Matching, MM)

有词典切分，机械切分

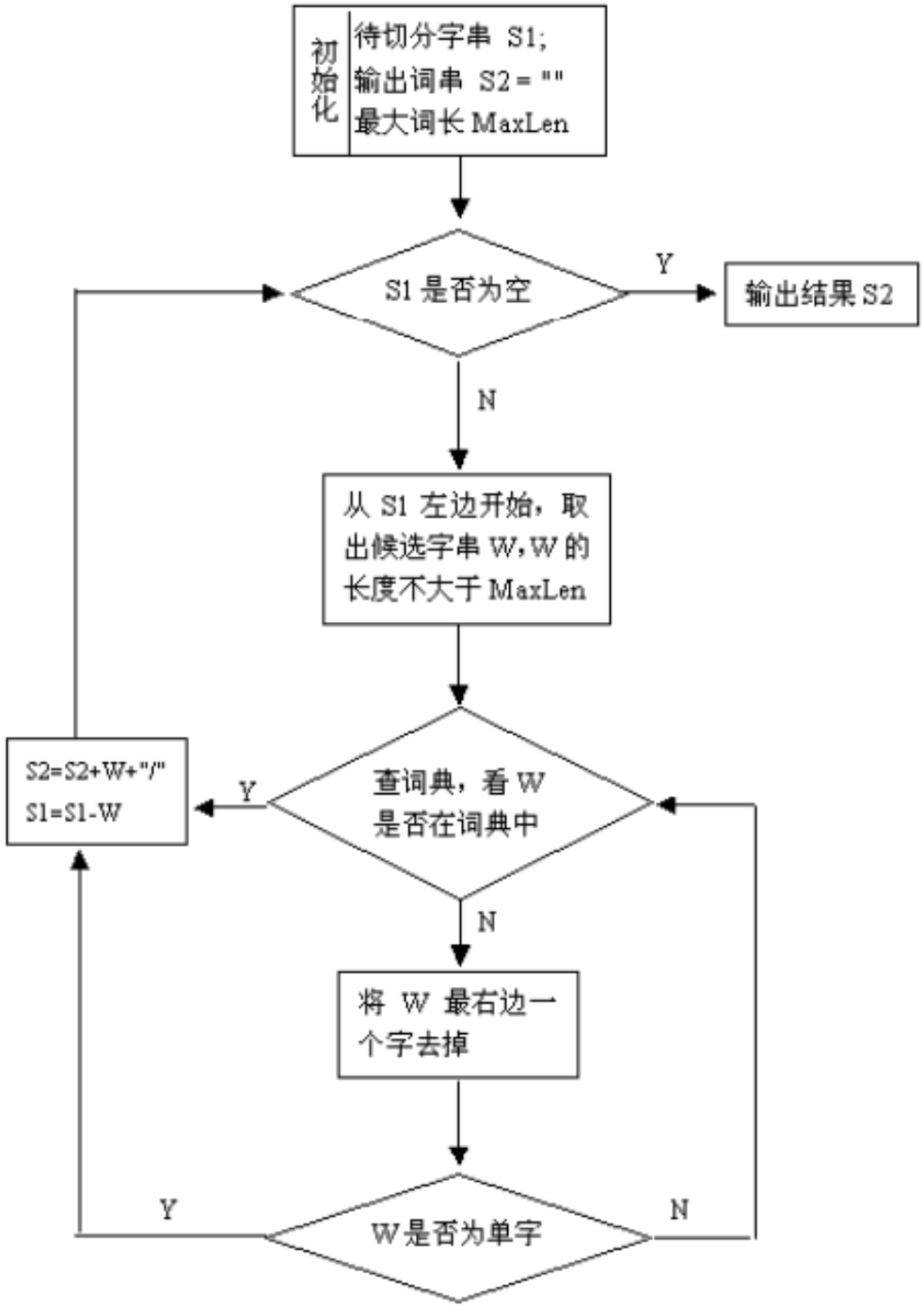
长词优先原则

正向最大匹配 **(Forward MM)**

逆向最大匹配 **(Backward MM)**

双向最大匹配算法

(Bi-directional MM)



最大匹配法分词示例 – 正向

词语
…
计算
语言
计算语言学
课程
课时
….

输入：S1=“计算语言学课程是两个课时"

输出：S2= " "

设定最大词长MaxLen = 5

(大规模真实语料中99%的词例的长度都在5字以内。

黄昌宁、赵海，2007，中文分词十年回顾，

《中文信息学报》2007年第3期，8-19页。)

W= “计算语言学”， S2=“计算语言学/”， S1=“课程是两个课时"

W=“课程是两个” … W=“课程是两” … W=“课程是” …

W=“课程”， S2=“计算语言学/课程/”， S1=“是两个课时"

W= “是两个课时” … W= “是两个课” …

W=“是” S2=“计算语言学/课程/是/” S1=“两个课时"

……

S1=“” S2=“计算语言学/课程/是/两/个/课时/”

最大匹配法分词示例 - 逆向

输入：S1=“计算语言学课程是两个课时"

输出：S2= “/ "

W= “是两个课时” … W= “两个课时” … W= “个课时” …

W= “课时”， S2=“课时/”，S1= “计算语言学课程是两个”

W=“课程是两个” … W=“程是两个” … W=“是两个” ……

W=“个”， S2=“个/课时/”， S1= “计算语言学课程是两”

W=“学课程是两” … …

W=“两”， S2=“两/个/课时/”， S1= “计算语言学课程是”

W=“言学课程是” ……

W=“是”，S2=“是/两/个/课时/”， S1= “计算语言学课程”

……

S1=“” S2=“计算语言学/课程/是/两/个/课时/”

词语
…
计算
语言
计算语言学
课程
课时
….

最大匹配法分词示例 – 双向

输入字串：他是研究生物化学的。

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/。

- 定位歧义 （用处）

最大匹配法

- 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；
- 缺点：
  - 歧义消解的能力差；
  - 切分正确率不高，一般在95％左右 （逆向略高于正向）。 有原因的：汉语句子的重心一般在右边

最大匹配法的问题

- 存在分词错误
  =〉增加知识，局部修改
- 无法发现分词歧义
  =〉从单向最大匹配改为双向最大匹配

  – 正向最大匹配和逆向最大匹配结果不同
    - **FMM　　有意/见/分歧/**
    - **BMM　　有/意见/分歧/**
  – 正向最大匹配和逆向最大匹配结果相同
    - **FMM & BMM　原子/结合/成分/子时/**

局部修改1：增加歧义词表，排歧规则

规则示例

IF W==“个人”, W_{left} == 数词, THEN W = "个/人/“ ENDIF

三/个/人

个人/英雄主义

歧义词表
……
才能
个人
家人
马上
研究所
……

局部修改2：增加“回溯”

- 对于某些交集型歧义，可以通过增加回溯机制来修改最大匹配法的分词结果。

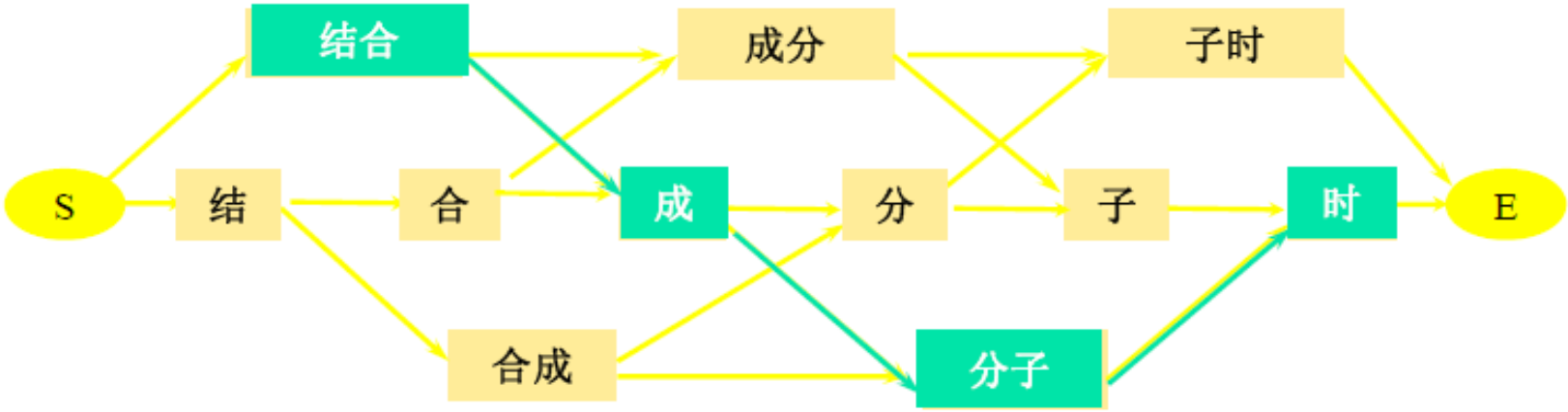
- 例如：“爱人民英雄”
  - 顺向扫描的结果是：“爱人/民/ 英雄/”
  - 通过查词典知道“民”不在词典中，于是进行回溯，将“爱人”的尾字“人”取出与后面的“民”组成“人民”
  - 再查 词典，看“爱”，“人民”是否在词典中，如果在，就将分词结果调整为：“爱/人民/英雄/”

最大匹配法分词的问题

- 双向最大匹配法可以发现链长为奇数的交集型歧义，但无法发现链长为偶数的交集型歧义
  - 链长为5： 鞭炮声响彻夜空
  - 链长为4： 中国产品质量
- 无法发现组合型歧义
  - 你/认为/学生会/ 听/老师/的/吗
- 在最大匹配法的基础上进行修改
  - 如何给出“改错”的触发条件带有一定的主观性
- 需要更全面地考虑分词的改进办法

汉语词语切分的数据结构—词图

例：结合成分子时



词图给出了一个字符串的全部切分可能性

分词任务：寻找一条起点S到终点E的最优路径

最短路径分词法

- *基本思想：*在词图上选择一条词数最少的路径
- 优点：好于单向的最大匹配方法
  - 最大匹配：独立自主/和平/等/互利/的/原则 **(6 words)**
  - 最短路径：独立自主/和/平等互利/的/原则 **(5 words)**
- 缺点：同样无法解决大部分交集型歧义
  - 结合/成分/子时
  - 他/说/的/确实/在理

    他/说/的确/实在/理　　（都是最短路径）

    他/说/的确/实/在理

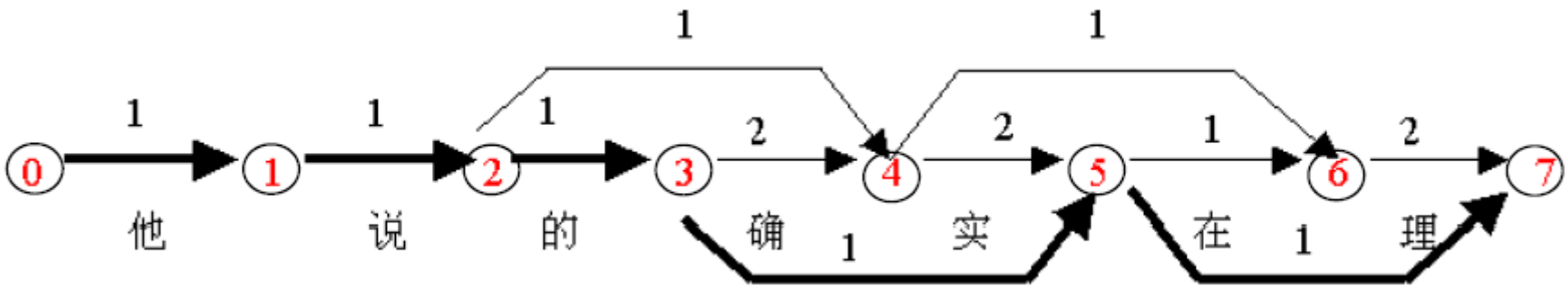
半词法分词

- *基本观察：*大多数单字在语境里如果能组成合适的词就不倾向于单独使用
- *基本概念：*
  - 半词：如果一个字不单独作为词使用，就是半词。半词既包含了成词语素，也包含了不成词语素，后者肯定是半词，比如“民”。前者则要看它作为语素的使用频度高，还是作为单字词的使用频度高，比如“见”。
  - 整词：如果一个字更倾向于自己成词，而不倾向于和别的字组成 词，这类“单字词”就称之为“整词”。这类词就是一般说的单字高频成词语素，比如“人、说、我”等。
- *基本思路：*充分利用半词和整词的差别，尽量选择没有半词落单的分词方案

半词法分词的实现

- 在词图的路径优劣评判中引入罚分机制
- 罚分规则：
  1. 每个词对应的边罚1分。
  2. 每个半词对应的边加罚1分。
  3. 一个分词方案的评分为它所对应的路径上所有边的罚分之和。
  4. 最优路径就是罚分最低的分词路径。

半词法分词示例



他/说/的/确实/在理 （1+1+1+1+1 = 5分）

他/说/的确/实/在理 （1+1+1+2+1 = 6分）

他/说/的确/实在/理 （1+1+1+1+2 = 6分）

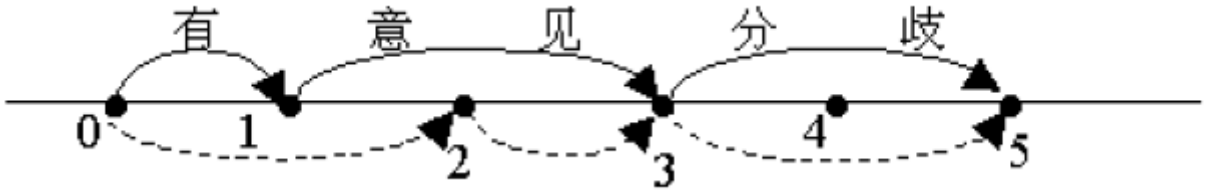
但是：仍然无法解决“有意见分歧”的问题

有/意见/分歧

有意/见/分歧

最大概率法分词

- *基本思想：*在词图上选择概率最大的分词路经作为最优结果



路径1：0-1-3-5　有/意见/分歧

路径2：0-2-3-5　有意/见/分歧

该走哪条路呢？

最大概率法分词

输入: 字符串**S:** 有意见分歧

输出: 词串 **W1:** 有/意见/分歧/

词串 **W2:** 有意/见/分歧/

Max(P(W1|S), P(W2|S))?

$$P(W|S) = \frac{P(S|W) \times P(W)}{P(S)} \approx P(W)$$

语言模型

Language Model

$$P(W) = P(w_1, w_2, \cdots, w_i)$$
$$= P(w_1) \times P(w_2|w_1) \times P(w_3|w_1 w_2) \times \cdots \times P(w_i|w_1, w_2, \cdots, w_{i-1})$$

最大概率法分词

1-gram，一元语法

$$P(W) \approx P(w_1) \times P(w_2|w_1) \times P(w_3|w_2) \cdots \times P(w_i|w_{i-1})$$

0-gram

$$P(W) \approx P(w_1) \times P(w_2) \times \cdots \times P(w_i)$$

独立性假设

$$P(w_i|w_{i-1}) = \frac{w_{i-1}w_i\text{在语料库中出现的次数}}{w_{i-1}\text{在语料库中出现的次数}} = \frac{Freq(w_{i-1}w_i)}{Freq(w_{i-1})}$$

$$P(w_i) = \frac{w_i\text{在语料库中出现的次数}}{\text{语料库中的总词数}N} = \frac{Freq(w_i)}{N}$$

最大概率法分词示例

输入：字符串S: 有意见分歧

输出：词串 W1: 有/意见/分歧/

词串 W2: 有意/见/分歧/

P(W1) = P(有) * P(意见) * P(分歧)

= 1.8 e-9

P(W2) = P(有意) * P(见) * P(分歧)

= 1.0 e-11

P(W1) > P(W2)

词语	概率
…	…
有	0.0180
有意	0.0005
意见	0.0010
见	0.0002
分歧	0.0001
…	…

最大概率法分词的实现
—如何高效求解最优路径？

- 动态规划算法：最优路径中的第 *i* 个词 W_i的累积概率等于它的左邻词W_{i-1}的累积概率乘以W_i自身的概率。

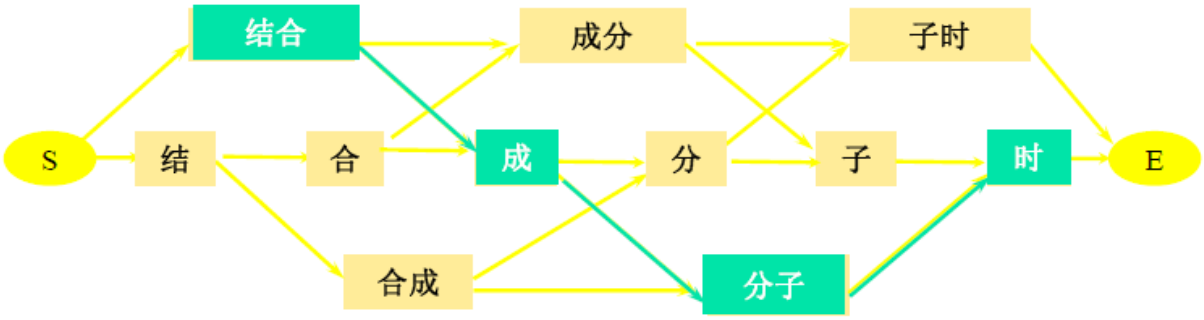
$$P'(w_i) = P'(w_{i-1}) \times P(w_i)$$

- 为方便计算，一般把概率转化为路径费用（代价）

$$C = -\log(P)$$

$$C'(w_i) = C'(w_{i-1}) + C(w_i) \qquad \text{公式1}$$

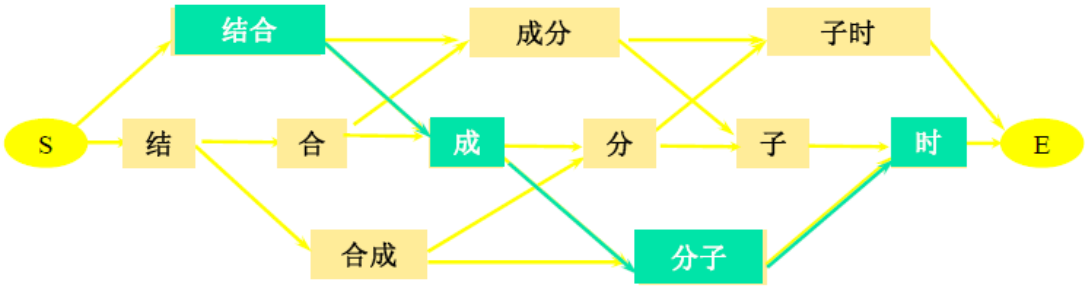
最小累积费用　　最佳左邻词



最大概率分词法的实现：动态规划算法(Dynamic Programming)

1. 对一个待分词的字串 S，按照从左到右的顺序取出全部候选词$w_1, w_2, \ldots, w_i, \ldots, w_n$；
2. 到词典中查出每个候选词的概率值$P(w_i)$，转换为费 用$C(w_i)$，并记录每个候选词的全部左邻词；
3. 按照公式1计算每个候选词的累计费用，同时比较得 到每个候选词的最佳左邻词；
4. 如果当前词w_n是字串S的尾词，且累计费用$C'(w_n)$最小，则w_n就是S的终点词；
5. 从w_n开始，按照从右到左顺序，依次将每个词的最佳左邻词输出，即为S的分词结果。

例：结合成分子时



序号	候选词	费用	累计费用	最佳左邻
0	结	3.573	3.573	-1
1	结合	3.543	3.543	-1
2	合	3.518	7.091	0
3	合成	4.194	7.767	0
4	成	2.800	6.343	1
5	成分	3.908	7.451	1
6	分	2.862	9.205	4
7	分子	3.465	9.808	4
8	子	3.304	10.755	5
9	子时	6.000	13.451	5
10	时	2.478	12.286	7

最大概率法分词的问题

- 优点：
  - 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率。

- 缺点：
  - 训练语料的规模和覆盖领域不好把握；
  - 计算量较大。

最大概率法分词的问题

- 并不能解决所有的交集型歧义问题
  - “这事的确定不下来”
    - **W1= 这/事/的确/定/不/下来/　　P(W1) < P(W2)**
    - **W2= 这/事/的/确定/不/下来/**
- 一般也无法解决组合型歧义问题
  - “做完作业才能看电视”
    - **W1= 做/完/作业/才能/看/电视/　P(W1) > P(W2)**
    - **W2= 做/完/作业/才/能/看/电视/**

由字构词（基于字标注）的分词方法

- 由字构词 (基于字标注)的分词方法发表在**2002**年第一届国际计算语言学学会**(ACL)**汉语特别兴趣小组**SIGHAN** 组织的汉语分词评测**(Bakeoff)**研讨会上。该方法在**2005**年和**2006**年的两次**Bakeoff** 评测中取得好成绩。

由字构词（基于字标注）的分词方法

- ***基本思想：***将分词过程看作是字的分类问题。该方法认为，每个字在构造一个特定的词语时都占据着 一个确定的构词位置(即词位)。假定每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。
- 这里所说的“字”不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在汉语文本中的文字符号，所有这些字符都是由字构词的基本单元。

基于字序列标注的方法

- 分词可以看做是对字加“词位标记”的过程
- “字”的词位分类：

B	E	M	S
词首	词尾	词中	独立词

自然句形式	已结婚的和尚未结婚的都应该到计生办登记
词切分结果	已/ 结婚/ 的/ 和/ 尚未/ 结婚/ 的/ 都/ 应该/ 到/ 计生办/ 登记/
字标注结果	已 结 婚 的 和 尚 未 结 婚 的 都 应 该 到 计 生 办 登 记 S B E S S B E B E S S B E S B M E B E

由字构词（基于字标注）的分词方法

1. 在字标注过程中，对所有的字根据预定义的特征进行词位特征学习，获得一个概率模型；
2. 然后在待切分字串上，根据字与字之间的结合紧密程度， 得到一个词位的分类结果；
3. 最后根据词位定义直接获得最终的分词结果。

- 工具：
  - 支持向量机(Support Vector Machine)
  - 条件随机场(Conditional Random Field)
  - 深度学习模型: **BiLSTM, RNN**

最常用的两类特征是字本身和词位(状态) 的转移概率

由字构词（基于字标注）的分词方法

- 该方法的重要优势在于，它能够平衡地看待词表词和未登录词的识别问题，文本中的词表词和未登录词都是用统一的字标注过程来实现的。

- 在学习构架上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词识别模块，因此，大大地简化了分词系统的设计。**[黄昌宁，2006]**

生成式方法与区分式方法的结合

- 大部分基于词的分词方法采用的是生成式模型
- 而基于字的分词方法采用区分式模型
- 基于字的区分模型有利于处理集外词，而基于词的生成模型更多地考虑了词汇之间以及词汇内部字与字之间的依存关系。因此，可以将两者的优势结合起来。

实验结果: 利用第二届 **SIGHAN Bakeoff** 评测语料**(2005)**

- **(1)** 基于词的 **3-gram: P＝89.8%**
- **(2)** 基于字的 **CRF: P＝94.3%**
- **(3)** 融合方法 **3-gram: P＝95.0%**

汉语自动分词

- **Python: jieba, pynlpir, SnowNLP, THULAC,… (GitHub)**
- **Urheen** 汉语自动分词系统：
  - **http://www.nlpr.ia.ac.cn/cip/software.htm**
- 其他分词方法
  - 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
  - ……

方法比较

- 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。
- 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。

方法比较

- 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。

- 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。

识别未登录词

不同类别未登录词识别难度的差异

- 较成熟
  - 中国人名、译名
  - 中国地名
- 较困难
  - 商标字号
  - 机构名
- 很困难
  - 专业术语
  - 缩略语
  - 新词语

占未登录 词的**95**％！

识别未登录词的策略

- 尽可能多地收集词汇，以降低碰到未登录词的机会
- 通过构词规则和上下文特征规则来识别
  - “雪村先生创作了很多歌曲”
- 通过统计的方法来猜测经过一般的分词过程后剩下的“连续单字词碎片”是人名、地名等的可能性，从而识别出未登录词
- 分而治之：对不同类的未登录词采用不同的办法识别

中国人名的内部构成规律 1

- 在汉语的未定义词中，中国人名是规律性最强，也是最容易识别的一类；
- 中国人名一般由以下部分组合而成：
  - 姓：张、王、李、刘、诸葛、西门、范徐丽泰
  - 名：李素丽，张华平，王杰，诸葛亮
  - 前缀：老王，小李
  - 后缀：王老，赵总
- 中国人名各组成部分用字比较有规律

中国人名的内部构成规律 2

- 根据统计，汉语姓氏大约有**1000**多个
  - 姓氏中使用频度最高的是“王”姓;
  - “王, 陈, 李, 张, 刘” 等**5**个大姓覆盖率达**32%;**
  - 姓氏频度表中的前**14**个高频度的姓氏覆盖率为**50%;**
  - 前**400**个姓氏覆盖率达**99%**。
- 人名的用字也比较集中
  - 频度最高的前**6**个字覆盖率达**10.35%;**
  - 前**10**个字的覆盖率达**14.936%;**
  - 前**15**个字的覆盖率达**19.695%;**
  - 前**400**个字的覆盖率达**90%**。
  - *男性：志，建，文，华，明，辉，伟*
  - *女性：丽，晓，华，英，玲，芳，秀*

中国人名的内部构成规律 3

中国人名内部各组成部分的组合规律

- 姓＋名
  - 姓、名均可再分“单字”、“双字”
- 前缀＋姓
- 姓＋后缀
- 姓＋姓＋名（海外已婚妇女）

中国人名的上下文构成规律

- 身份词：
  - 前：工人、教师、影星、犯人
  - 后：先生、同志
  - 前后：校长、经理、主任、医生
- 地名或机构名的字结构：
  - 前：静海县大丘庄禹作敏
  - 前：年过七旬的王贵芝
- 动作词
  - 前：批评、逮捕、选举
  - 后：说、表示、吃、结婚
- ……

中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
  - 姓氏：于，马，黄，张，向，常，高
  - 名字：周鹏和同学，周鹏和同学
- 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
  - [王国]维、[高峰]、[汪洋]、张[朝阳]、冯[胜利]
- 人名与其上下文组合成词
  - 这里[有关]天培的壮烈
  - 费孝[通向]人大常委会提交书面报告
  - 邓颖[超生]前使用过的物品
  - 赵[微笑]着走了
- 人名地名冲突
  - 河北省刘庄

中国人名识别的方法

- 不考虑上下文信息的识别方法：
  - 根据一个汉字串内部各汉字的特征计算该汉字串作为中文姓名的概率
  - ***Sproat R. et al., 1996, A Stochastic Finite-state Word Segmentation Algorithm for Chinese, Computational Linguistics, Vol.22, No.3, pp377-404.***
- 考虑上下文信息的识别方法：
  - 把姓名及其上下文中的汉字标记不同的“角色”，将人名识别问题转换为汉字的角色标注问题
  - *张华平、刘群，**2004**，基于角色标注的中国人名自动识别研究，《计算机学报》**Vol.27, No.1***

中国人名识别的方法

- 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阈值(threshold value)，根据姓名概率和修饰规则对潜在的姓名进行筛选。

中国人名识别的方法

计算概率估计值

- 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：

$$F(X) = \frac{X\text{用做姓氏的次数}}{\text{X出现的总次数}}$$

$$F(m_1) = \frac{m_1\text{用做名字首字的次数}}{m_1\text{出现的总次数}}$$

$$F(m_2) = \frac{m_2\text{用做名字尾字的次数}}{m_2\text{出现的总次数}}$$

*Cname*做为姓名的概率：

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

中国人名识别的方法

- 使用修饰规则：
  - 如果姓名前是一个数字，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。
  - 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。

音译名的识别 1

- 音译名用字非常集中《英语姓名译名手册》中共收英语姓氏, 教名约4万个, 经计算机统计得出英语姓名译名用字表共476个：
- “啊阿埃艾爱昂奥巴白柏拜班邦包保堡鲍北贝倍本比彼边别滨宾玻波博勃伯卜布采蔡藏策查察 昌彻陈楚垂茨慈次聪存措达大戴代丹当道德得登邓迪底地蒂第帝丁东杜敦顿多厄恩耳尔法凡范 方菲费芬丰冯佛夫福弗辅富盖甘冈高哥戈葛格各根贡古顾瓜圭郭果哈海罕翰汉杭豪赫黑亨洪侯 胡华怀惠霍基吉季计嘉佳加贾简姜焦杰捷金津京久居喀卡开凯坎康考柯科可克肯孔扣寇库夸匡 奎魁坤昆阔拉腊莱来赖兰朗劳勒乐雷黎理李里礼荔丽历利立莲连廉良列琳林霖龄留刘流柳龙隆 卢鲁露路吕略伦萝罗洛玛马麦迈满曼芒茅梅门蒙孟米密敏明名摩莫墨默姆木穆拿娜纳乃奈南内 嫩能妮尼年涅宁牛纽农努女诺欧帕派潘庞培佩彭蓬皮匹平泼朴普漆奇齐契恰钱强乔切钦琴青琼 丘邱屈让热仁日荣茹儒瑞若撒萨塞赛三缮桑瑟森莎沙珊山尚绍舍申生盛圣施诗石什史士寿舒朔 斯思丝松孙索所塔泰坦汤唐陶特藤提惕田铁汀廷亭通透图托脱娃瓦万旺威韦为维伟魏卫温文翁 沃乌武伍西锡希悉席霞夏显香向晓肖歇谢欣辛兴幸姓雄休修雪逊雅亚延扬阳尧耀耶叶依易意因 英永尤雨约宰赞早泽曾扎詹湛章张哲者珍真芝知智治朱卓兹子宗祖佐丕谟葆薇岑弼娅缪珀瑙赉 滕斐熙鸠窦艮麟黛”。

辛华编《英语姓名译名手册》商务印书馆1973年（修订版）

新华通讯社译名资料组编《英语姓名译名手册》商务印书馆1997年（第二次修订版）

音译名的识别 2

- 音译名内部很难划分出结构，但有一些常见音节，如“斯基、斯坦”等
- 不同语言的音译规律不尽相同，如法语、俄语、蒙古语译名用字与英语就有较大区别（蒙语人名举例：“那顺乌日图、青格勒图”），如果按不同的语言训练不同的模型可能会比使用统一的模型效果更好
- 音译名可以是人名、地名或其他专名，上下文规律差别较大
- 由于音译名用字比较集中，识别正确率较高

中国地名的识别

- 中国地名委员会编写了《中华人民共和国地名录》，收集了全国乡镇以上（含乡镇）各级行政区域的名称，以乡镇人民政府所在地为主的居民聚落名称，山、河、湖、海、岛、高原、盆地、沙溪等自然地理实体名称，名胜古迹、纪念地、古遗址、水库、桥梁、电站等名称。
- 共收录地名**10**万多条。这个地名录中使用的汉字共**2662**个，频度最高的前**65**个汉字占总频度的**50.22%**，前**622**个汉字占总频度的**90.01%**，前**1872**个汉字占总频度的**99%**。
- 与人名的用字情况相比较，地名用字分散得多。
- 地名内部也有一定的结构，右边界比左边界更容易识别。

中国地名的识别

- 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
- 基本方法
  - 统计模型
  - 通过训练语料选取阈值
  - 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

机构名的内部构成规律 1

- 机构名一般都是定中结构（如：教育部语信司）
- 机构名的后缀一般比较集中，识别相对容易
- 机构名左边界识别非常困难
- 机构名中含有大量的人名、地名、企业字号等专有名称。在这些专有名称中，地名所占的比例最大，其中未登录地名又占了相当一部分的比例。所以机构名识别应在人名、地名等其他专名识别之后进行，其他专名识别的正确率对机构名识别正确率有较大影响

机构名的内部构成规律 2

- 中文机构名用词非常广泛: 通过对人民日报1998年1月中的10817个机构名所含的19986个词进行统计，共计27种词，其中名词最多（9941个），地名其次（5023个），以下依次为简称（1169个）、专有名词（1125个）、动词（848个）以及机构名（714个）等
- 机构名长度极其不固定
- 机构名很不稳定: 随着社会发展，新机构不断涌现，旧机构不断被淘汰、改组或更名

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

4　对文本分词质量的评价

- 计算分词正确率的不同标准
  - 以词数算
  - 以句数算

- 分词质量对**NLP**应用系统的影响 **(**“移动电话”**)**
  - 分词质量对**MT**的影响　从合 —— 对翻译、校对有利
  - 分词质量对**IR**的影响　从分 —— 对**IR**有利
  - **……**

准确率、召回率、F-measure

- 准确率 **(precision)**

$$Precision(P) = \frac{\text{分词结果中正确分词数}}{\text{结果中所有分词数}} \times 100\%$$

- 召回率 **(recall)**

$$Recall(R) = \frac{\text{分词结果中正确分词数}}{\text{标准答案中所有分词数}} \times 100\%$$

- **F-**评价**(F-measure** 综合准确率和召回率的评价指标**)**

$$F - measure = \frac{2PR}{P + R}$$

中文分词效果评测

- 国内**863**计划，**973**计划，中文信息学会组织过多次评测
- 国际上**SIGHAN bakeoff 2003 –2007** **http://www.sighan.org/**

历届Sighan 在City-U 语料上评测结果F 值最好成绩

	Recall	Precision	F-score	Roov	训练语料词数	测试语料词数
2007	0.9526	0.9493	0.9510	0.7495	1.04M/43K	230K/23K
2006	0.9730	0.9720	0.9720	0.7870	1.6M/76K	220K/23K
2005	0.9410	0.9460	0.9430	0.6980	1.46M/69K	41K/9K
2003	0.9470	0.9340	0.9400	0.6250	240K	35K

- 刘群、钱跃良，2008，中文信息处理技术评测综述，《中国计算机学会通讯》2008 年第2期。

小 结

- 英语单词识别和还原
- 中文分词的问题
  - 词的界定：规范+词表+语料库
  - 歧义：交集型、组合型、链长
  - 未登录词：类型
- 分词的方法
  - 从句到词: (a)最大匹配 (b)最佳路径
  - 从字到词：字序列标注法
- 未登录词的识别
- 分词的评价
  - 指标：(1) 准确率P (2) 召回率R (3) F-Score
  - 评测：SIGHAN bakeoff等