中文自动分词

苏州大学计算机科学与技术学院

问题的提出

- 例子
  - He will come to Shanghai tomorrow (英文)
  - 他明天将来上海 (中文)
- 区别
  - 中文中最小单位是字，英文为单词(有意义)
  - 中文中具有意义的最小单位是词(含单字词)
  - 中文中字与字或词与词之间没有明显的界限

英语词语切分问题

- 英语中不是完全没有词语切分问题
- 不能仅凭借空格和标点符号切分
  - 缩写词

    N.A.T.O　i.e.　m.p.h　Mr.　AT&T
  - 连写形式以及所有格结尾

    I'm　He'd　don't　Tom's
  - 数字、日期、编号

    128,236　+32.56　-40.23　02/02/94　02-02-94
  - 带连字符的词

    text-to-speech　text-based　e-mail　co-operate
- 英语中的切分通常被叫做**Tokenization**
- 和中文相比，英语切分问题较为容易

主要内容

- **基本概念**
- 分词规范
- 分词词典
- 分词算法
- 分词系统

基本概念

- 分词
  - 将连续的字串或字符序列按照一定的规范重新组合成词序列的过程
- 中文分词
  - 把中文的序列切分成有意义的词
  - 例：他/ 明天/ 将/ 来/ 上海/

意义与应用

- 研究意义
  - 中文信息处理的基础步骤
  - 对中文处理系统的性能有重要影响
  - 广泛的应用价值：语音识别、信息检索……
- 应用
  - 同音字、多音字识别
  - 文本校对
    - 抛妻别字 —— 抛弃别字 （字音编码输入）
    - 于预 —— 干预 （字形编码输入）
  - 简繁转换
    - 後面，皇后 —— 后
    - 松树，鬆开 —— 松

中文分词歧义

- 白痴造句法
  - 难过
    - 我家门前的小河很难过
  - 白痴
    - 小白痴痴地在门前等小黑回来
  - 如果
    - 汽水不如果汁好喝
  - 本来
    - 拿书本来打头会很痛

中文分词关键问题

- 切分歧义消解
  - 存在多种不同切分方法
    - 例：汽水不如果汁好喝
    - 汽水/不如/果汁/好喝
    - 汽水/不/如果/汁/好喝
- 未登录词识别
  - 词典中没有的词
    - 新的概念：高富帅 中国大妈
    - 新的专用名词：郭美美

歧义例子

- 这个学生会打蓝球
  - 这个/学生/会/打/篮球
  - 这个/学生会/打/篮球
- 你认为学生会听老师的吗
  - 你/认为/学生会/听/老师/的/吗
  - 你/认为/学生/会/听/老师/的/吗
- 南京市长江大桥
  - 南京市/长江大桥
  - 南京/市长/江大桥

歧义例子续

- 当结合成分子时
  - 当/结合/成分/子时
  - 当/结合/成/分子/时
  - 当/结/合成/分子/时
  - 当/结/合成分/子时

切分歧义

- 交集型歧义
  - 如果AB和BC都是词典中的词，那么如果待切分字串中包含“ABC”这个子串，就必然会造成两种可能的切分
    - “AB/ C/ ” 和 “A/ BC/ ”
  - 例：
    - 结合/成，结/合成
    - 网球/ 场，网/ 球场

切分歧义

- 组合型歧义
  - 如果AB和A、B都是词典中的词，那么如果待切分字串中包含“AB”这个子串，就必然会造成两种可能的切分
    - “AB/ ” 和“A/ B/ ”
  - 例：
    - 门/把手/坏/了，请/把/手/拿/开
    - 我/现/在/北京，我/现在/去/北京

切分歧义

- 混合型歧义
  - 同时包含交集型歧义和组合型歧义
  - 例：　这样的/人/才能/经受住考验
    这样的/人才/能/经受住考验
    这样的/人/才/能/经受住考验

中文文本中，交集型歧义与组合型歧义出现的比例约为1:22

切分歧义

真歧义

- 歧义字段在不同的语境中确实有多种切分形式
- 例：　　地面积

伪歧义	94%	这块/地/面积/还真不小		
真歧义	6%	地面/积/了厚厚的雪		
		多种切分形式均匀分布	0.72%	将技术/应用/于/项目　　精力/应/用于/学习
		一种切分形式占优	5.28%	解除/了/职务　　方程的/解/除了/0还有1

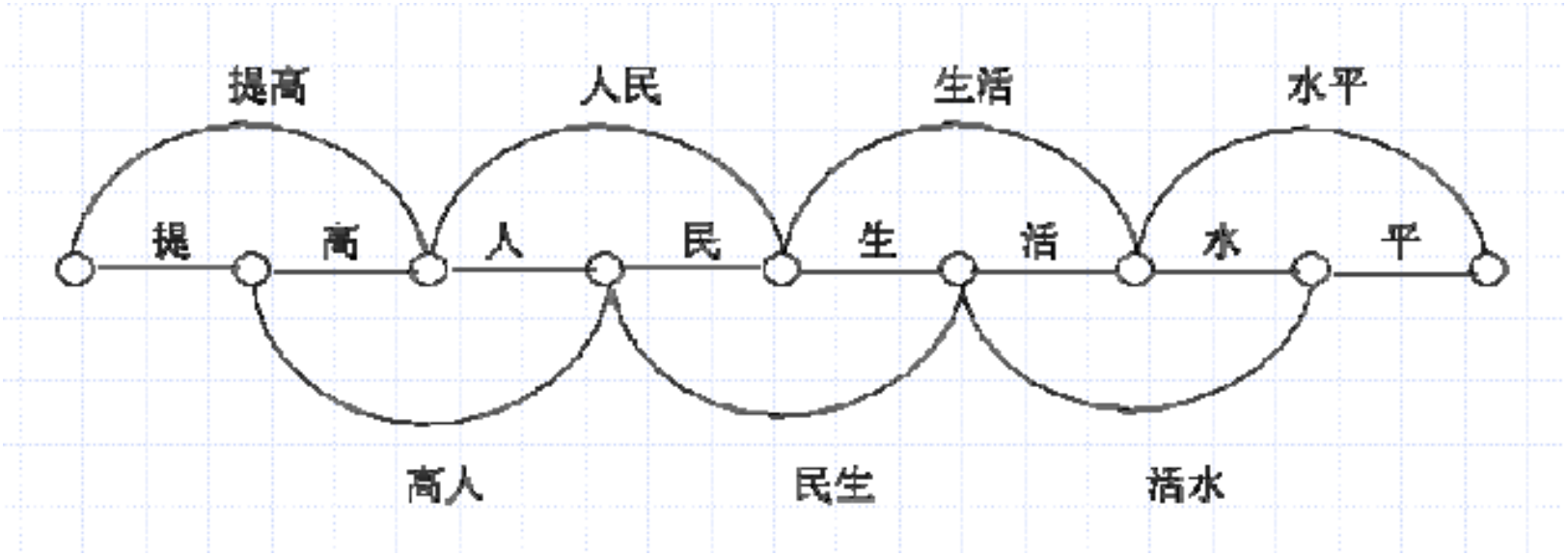
伪歧义

- 歧义字段单独拿出来看有歧义，但在所有真实语境中，仅有一种切分形式可接受
- 例：　　挨批评

  挨/批评（√）　　挨批/评（×）

- 对于交集型歧义字段，真实文本中伪歧义现象远多于真歧义现象

歧义切分的表示



未登录词

- 实体名词和专有名词
  - 中国人名: 王莉 张杰 李晓静
  - 中国地名：苏州 上海 百步街
  - 翻译人名：奥巴马 普京 莎拉波娃
  - 翻译地名：芝加哥 阿姆斯特丹 对马海峡
  - 机构名　：苏州大学　苹果公司　联合国　科技部
  - 商标字号：茅台 东阿阿胶 可口可乐
- 专业术语和新词语
  - 专业术语：无线应用通信标准协议 普适计算
  - 缩略语　：苏大 计科
  - 新词语　：裸官 程序猿

影响比较

- 国际中文自然语言处理评测
  - 每届Bakeoff都对每个语料库进行带有未登录词的基线和不含未登录词的顶线两种切分。F_{base}和F_{top}分别表示基线和顶线的**F**值
  - F_{top}-F_{base}表示未登录词带来的精度失落
  - 1-F_{top}表示切分歧义造成的精度失落
  - Bakeoff 2003和Bakeoff 2005语料库统计：
    - 精度失落：未登录词是切分歧义的**5.6-25.6**倍

未登录词识别

- 未登录词识别困难
  - 未登录词没有明确边界
    例：张掖市民乐县
  - 许多未登录词的构成单元本身可以独立成词
    例：张建国　内塔尼亚胡说
  - 与普通词相似
    例：爱子面容俨然是父亲的“女性版”
  - 呈现一定的句法结构
    例：好又多、我爱我家房地产经纪公司
- 通常每一类未登录词都要构造专门的识别算法
- 识别依据
  - 内部构成规律（用字规律）
  - 外部环境（上下文）

未登录词识别现状

- 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

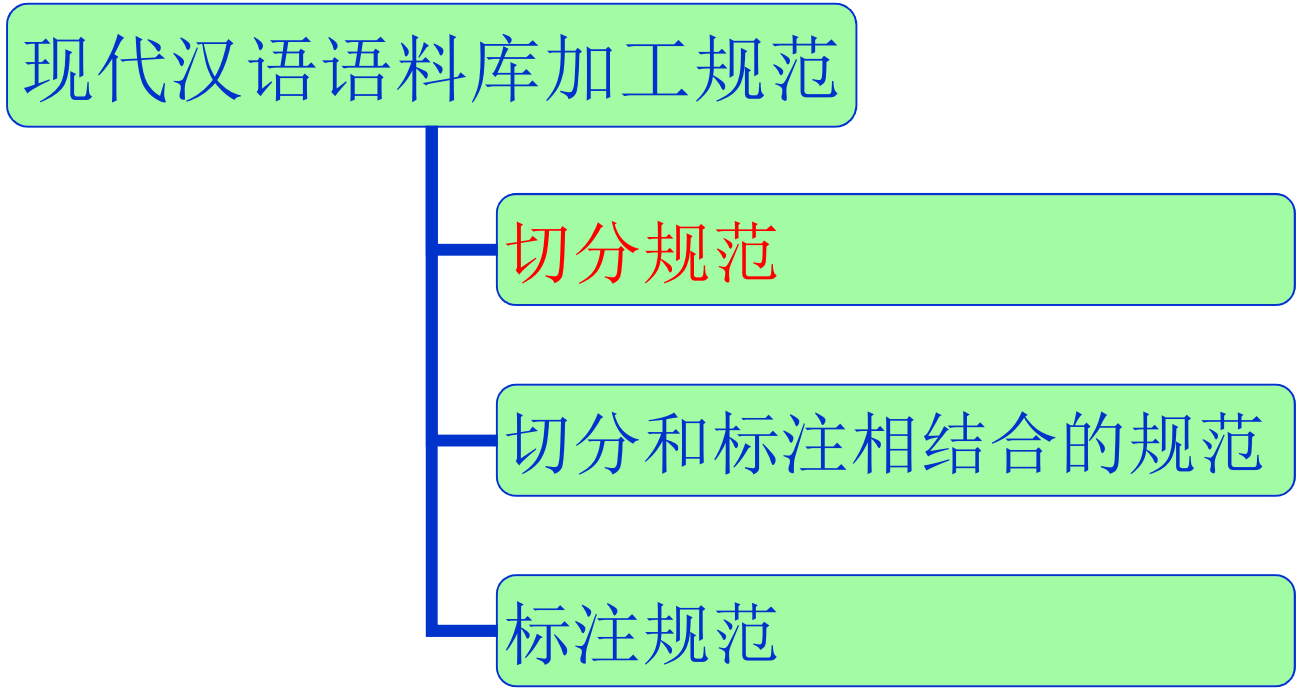
主要内容

- 基本概念
- **分词规范**
- 分词词典
- 分词算法
- 分词系统

分词规范

- 《信息处理用现代汉语分词规范》GB13715 1993
- 《资讯处理用中文分词规范》“台湾中研院” 1995
- 《现代汉语语料库加工规范——词语切分与词性标注》北大计算语言研究所 俞士汶等 1999

分词规范



分词规范

- 《现代汉语语料库加工规范》
  - 切分单位：沿用“分词单位”，主要是**词**，也包括了一部分结合紧密、使用稳定的**词组**。在某些特殊情况下孤立的语素或非语素字也可能出现在切分序列中。

    例：出/v 了/u 一/m 次/q 差/Ng
  - 人名：一些著名作者的或不易区分姓和名的笔名通常作为一个切分单位。

    例：鲁迅/nr， 巴金/nr， 琼瑶/nr

分词规范

- 《现代汉语语料库加工规范》
  - 地名：后有“省”、“市” 等单字的行政区划名称时，不切分。

    例：江苏省/ns， 上海市/ns
  - 数量词：切分为数词和量词，但少数数量词已是词典的登录单位，则不再切分。

    例：三/m 个/q，10/m 公斤/q，一个/m

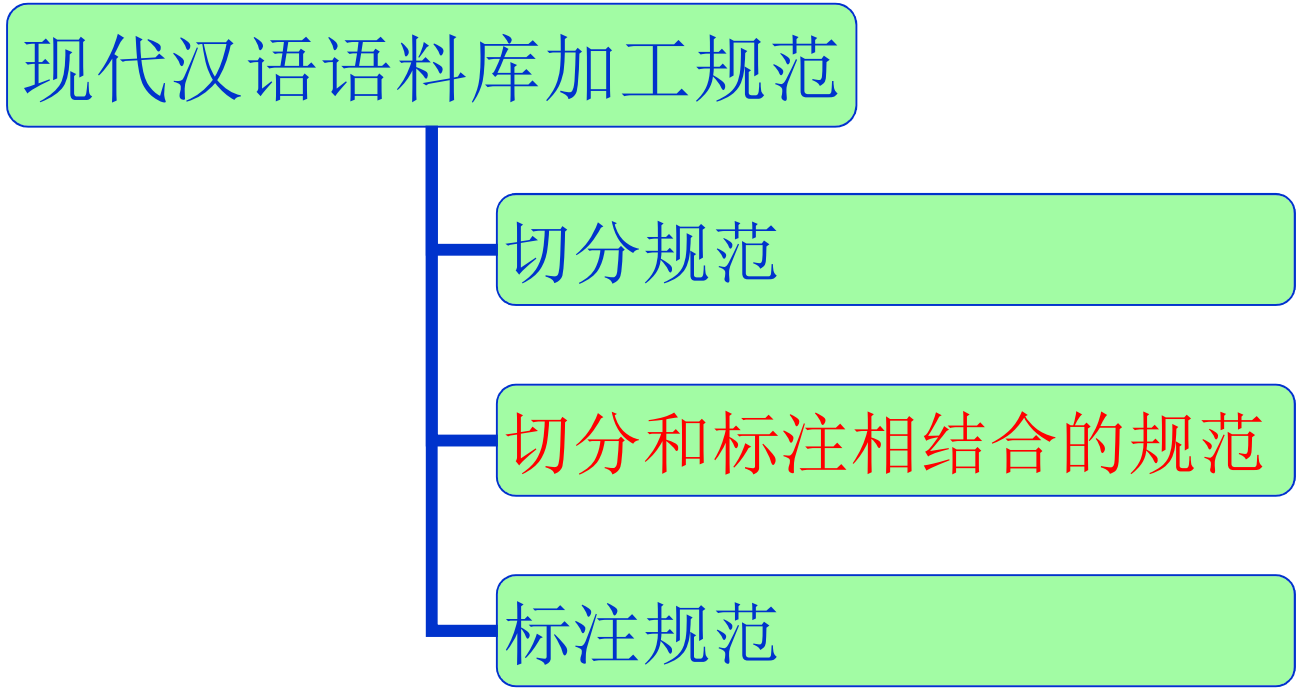
分词规范（续4）

- 《现代汉语语料库加工规范》
  - 单音节代词：“本”、“每”、“各”、“诸”后接单音节名词时，与其合为代词；当后接双音节名词时，应予切分。

    例：本报/r，　每人/r，本/r 地区/n
  - 四字的成语或习惯用语为一个切分单位。

    例：胸有成竹/i，由此可见/l
  - ……

分词规范（续5）



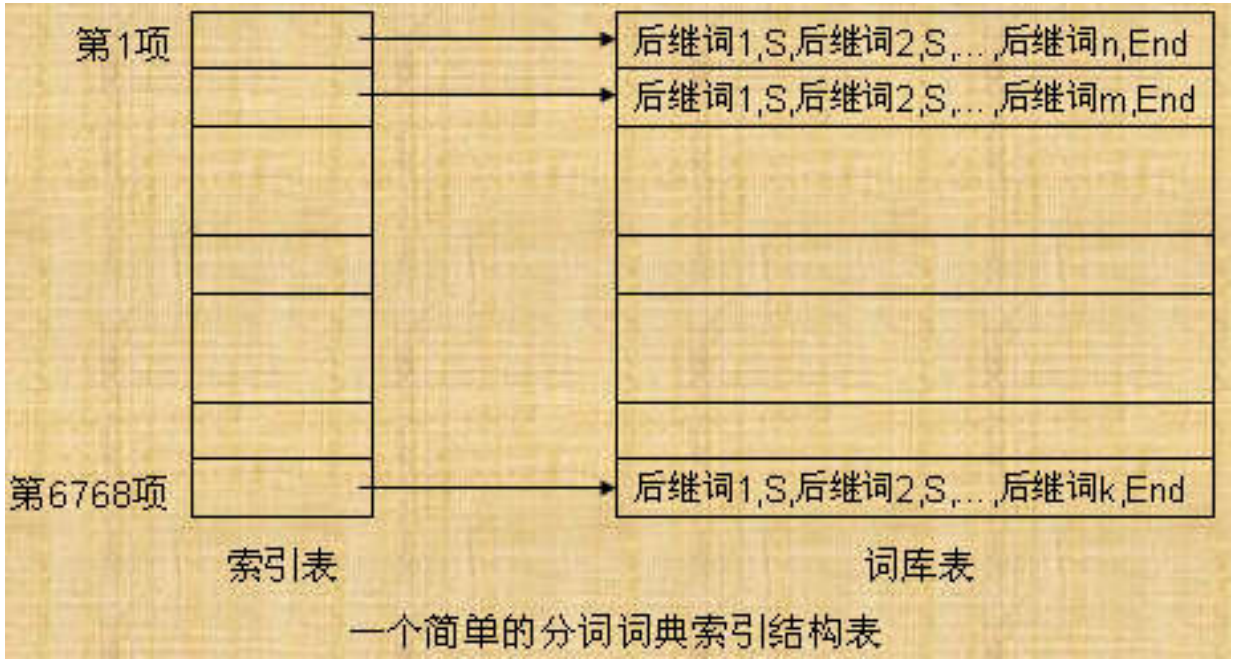
分词规范（续6）

- 《现代汉语语料库加工规范》
  - 重叠：
    - AA，AA的/地，AAB，ABB，AABB，A 里AB，A 不AB等形式不切分，ABAB形式切分
    - 人人/n，甜甜的/z，挥挥手/v，亮堂堂/z，方方面面/n，糊里糊涂/z，高兴/a 高兴/a
  - 附加：
    - 前接成分＋语素或词、语素或词＋后接成分、前接成分＋语素或词＋后接成分不切分
    - 阿华/nr，老张/nr，花儿/n，爷儿们/n，求知者/n，无政府主义者/n
  - 复合：
    - 如单纯方位词＋名（单音）的定中结构作为一个切分单位
    - 前院/s，左肩/n，后天/t

主要内容

- 基本概念
- 分词规范
- **分词词典**
- 分词算法
- 分词系统

词典构造



一个简单的分词词典索引结构表

—— 用于基于词典的分词算法

考虑因素

- 构建一个高效的分词词典，从以下三个方面考虑：
  - 查询速度：匹配算法效率的直接决定因素
  - 存储利用率：分词词典设计小
  - 维护效能：插入、删除、更新等操作的难易程度

词典举例

- 文本形式与数据库形式

阿爸
阿保之功
阿保之劳
阿波罗
阿波罗神
阿伯
阿爹
阿斗
阿尔巴尼亚
阿尔卑斯

词语代号	词语	出现次数
912	奥委会	110
913	奥希金斯	1
914	奥新社	1
915	奥运	46
916	奥运村	9
917	奥运会	241
918	澳	290
919	澳大利队	1
920	澳大利亚	364

主要内容

- 基本概念
- 分词规范
- 分词词典
- **分词算法**
- 分词系统

分词算法

- 正/逆向最大匹配
- 正/逆向最小匹配
- 邻近匹配
- 最短路径匹配
- 基于统计的算法

正/逆向最大匹配

- 正向最大匹配
  - 用MAXL表示最大词长，按照从左到右的顺序，首先从汉字串中取长度为MAXL的子串查词典。若词典中存在这个词，则切分出该子串，指针后移 MAXL 个汉字后继续切分，否则，子串长度减一，再与词典匹配。若长度为2的子串还不能在词典中查到，则取当前汉字为词，指针后移一个汉字继续匹配。
- 逆向最大匹配
  - 与前者区别在于抽取顺序，从汉字串尾端开始抽取。

正/逆向最大匹配（例）

- “他们明天来上海”(MAXL=4)
  - 正向最大匹配过程

    他们明天　他们明　他们　明天来上　明天来

    明天　来上海　来上　来　上海
  - 逆向最大匹配过程

    天来上海　来上海　上海　们明天来　明天来

    天来　来　他们明天　们明天　明天　他们

两个最大匹配

- 正向最大匹配
  - Forward Maximum Matching method, FMM
  - 错误切分率为1／169
  - 对交叉歧义和组合歧义没有什么好的解决办法
- 逆向最大匹配
  - Backward Maximum Matching method, BMM
  - 错误切分率为1／245
- 双向匹配法
  - Bi-direction Matching method, BM
  - 可以识别出分词中的交叉歧义

分词算法

- 正/逆向最大匹配
- 正/逆向最小匹配
- 邻近匹配
- 最短路径匹配
- 基于统计的算法

正/逆向最小匹配

- 正向最小匹配
  - 按照从左到右的顺序，首先从汉字串中取长度为2的子串查词典。若词典中存在这个词，则切分出该子串，指针后移2个汉字，否则，子串长度逐次加一继续匹配。若一直到长度为MAXL的子串仍无法匹配，则切分出当前汉字。
- 逆向最小匹配
  - 与前者区别在于抽取顺序，从汉字串尾端开始抽取。

正/逆向最小匹配（例）

- “他们明天来上海”
  - 正向最小匹配过程

    他们 明天 来上 来上海 来 上海
  - 逆向最小匹配过程

    上海 天来 明天来 们明天来 来 明天 他们

分词算法

- 正/逆向最大匹配
- 正/逆向最小匹配
- 邻近匹配
- 最短路径匹配
- 基于统计的算法

邻近匹配

- 算法
  - 设待切分中文字串 $C_0C_1C_2\cdots\cdots C_{n-1}$
  - 根据C_0C_1得到所有以C_0C_1为首的词条集W
  - 如果W为空，则将C_0切分出来
  - 否则切出满足： $\max\{k \mid C_0C_1\cdots\cdots C_k \in W\}$ 的子串 $C_0C_1\cdots\cdots C_k$
  - 再将剩余字串 $C_{k+1}C_{k+2}\cdots\cdots C_{n-1}$ 作为新的待切分串进行同样的处理
  - 直到待切分串变成空为止
- 改进的正向最大匹配，以降低时间复杂度。

邻近匹配（例）

- “为奥运会健儿加油啊”

“为奥” W = Ø 为

“奥运” W = {奥运，奥运会 }

“奥运”匹配 长度=2

“奥运会”匹配 长度=3 奥运会

“健儿” W = {健儿}

“健儿”匹配 长度=2 健儿

“加油” W = {加油，加油站}

“加油”匹配 长度=2

“加油站”≠“加油啊”不匹配 加油

“啊” 啊

分词算法

- 正/逆向最大匹配
- 正/逆向最小匹配
- 邻近匹配
- 最短路径匹配
- 基于统计的算法

最短路径匹配

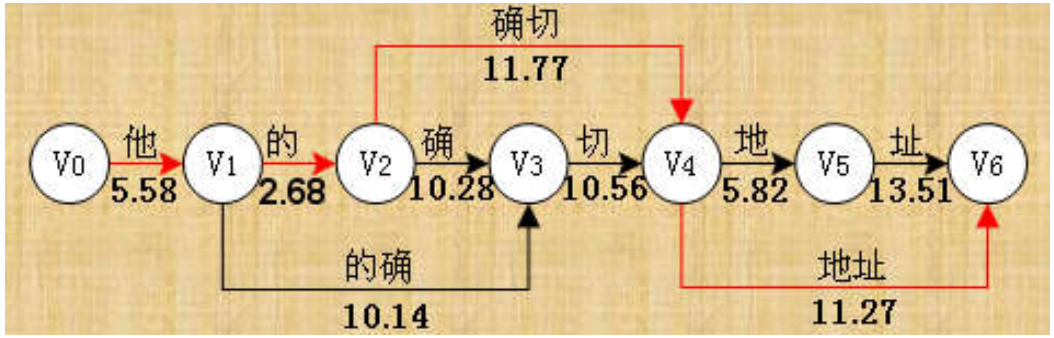
- 算法
  - 设待分中文字串$C_1C_1\ldots C_n$
  - 建立一个结点数为n+1的切分有向无环图G
  - 各结点编号依次为V_0，$V_1\ldots V_n$，通过以下两种方式建立G所有可能的词边
  - (1) 相邻结点V_{k-1}，V_k之间建立有向边$\langle V_{k-1}, V_k\rangle$，对应的词为$C_k$，边的权值$L_k=\ln(K)-\ln(K_i)$，其中K为词典所有词的频次之和，$K_i$为$C_k$出现的频次
  - (2) 若$w=C_iC_{i+1}\ldots C_j$是字典中词，则结点V_{i-1}，V_j之间建立有向边$\langle V_{i-1}, V_j\rangle$，对应的词为w，边的权值$L_w=\ln(K)-\ln(K_w)$，其中$K_w$为w出现的频次。

最短路径匹配（续）

- 采用Dijkstra算法求有向图G的最短路径
  - 如：路径为V_0-V_i-V_s-…-V_t-V_j-V_n，则词串切分为C_1…C_i/ C_{i+1}…C_s/ ……/ C_{t+1}…C_j/ C_{j+1}…C_n。
- 一般的最短路径匹配算法，将词对应边长的权值均设为1，但往往存在多条最短路径，如果只保留其中一个结果，切分效果不是很理想。

最短路径匹配（例）

- “他的确切地址”



词典片断

词语代号	词语	出现次数
15718	的	355225
15725	的确	206
16235	地	15366
16482	地址	66
62748	切	134
64892	确	179
64902	确切	40
76188	他	19618
103436	址	7

注：词典词频总和：5196663

最短路径：V_0-V_1-V_2-V_4-V_6

对应词串：C_1/ C_2/ C_3C_4/ C_5C_6

切分结果：他/ 的/ 确切/ 地址/

分词算法

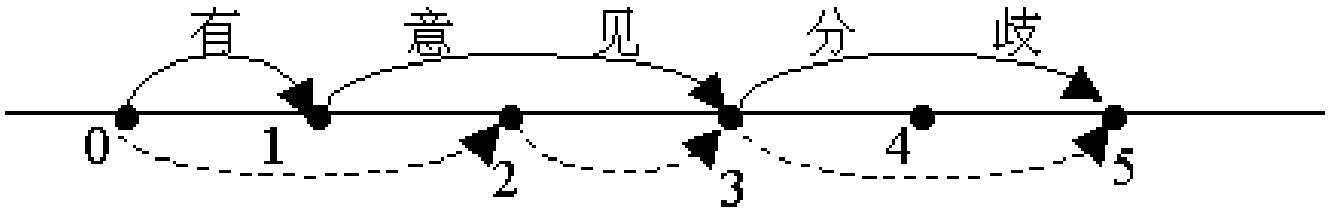
- 正/逆向最大匹配
- 正/逆向最小匹配
- 邻近匹配
- 最短路径匹配
- 基于统计的算法

最大概率法分词

基本思想是：

（1）一个待切分的汉字串可能包含多种分词结果

（2）将其中概率最大的那个作为该字串的分词结果



路径1： 0－1－3－5

路径2： 0－2－3－5

该走哪条路呢？

最大概率法分词

- S: 有意见分歧
  - W1: 有/ 意见/ 分歧/
  - W2: 有意/ 见/ 分歧/

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

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

$$P(W) = P(w_1, w_2, ..., w_i) \approx P(w_1) \times P(w_2) \times ... \times P(w_i)$$

独立性假设，一元语法

$$P(w_i) = \frac{w_i\text{在语料库中的出现次数}n}{\text{语料库中的总词数}N}$$

最大概率法分词

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

$P(W1) = P(有) * P(意见) * P(分歧)$

$= 1.8 \times 10^{-9}$

$P(W2) = P(有意) * P(见) * P(分歧)$

$= 1 \times 10^{-11}$

$P(W1) > P(W2)$

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/

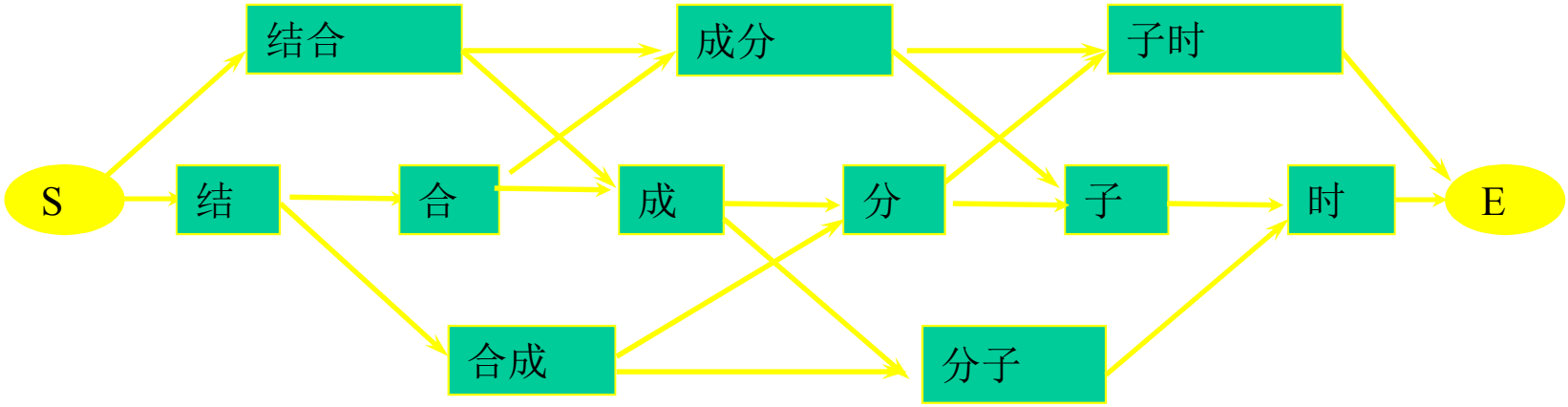
- 无法解决组合型歧义问题

  “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/

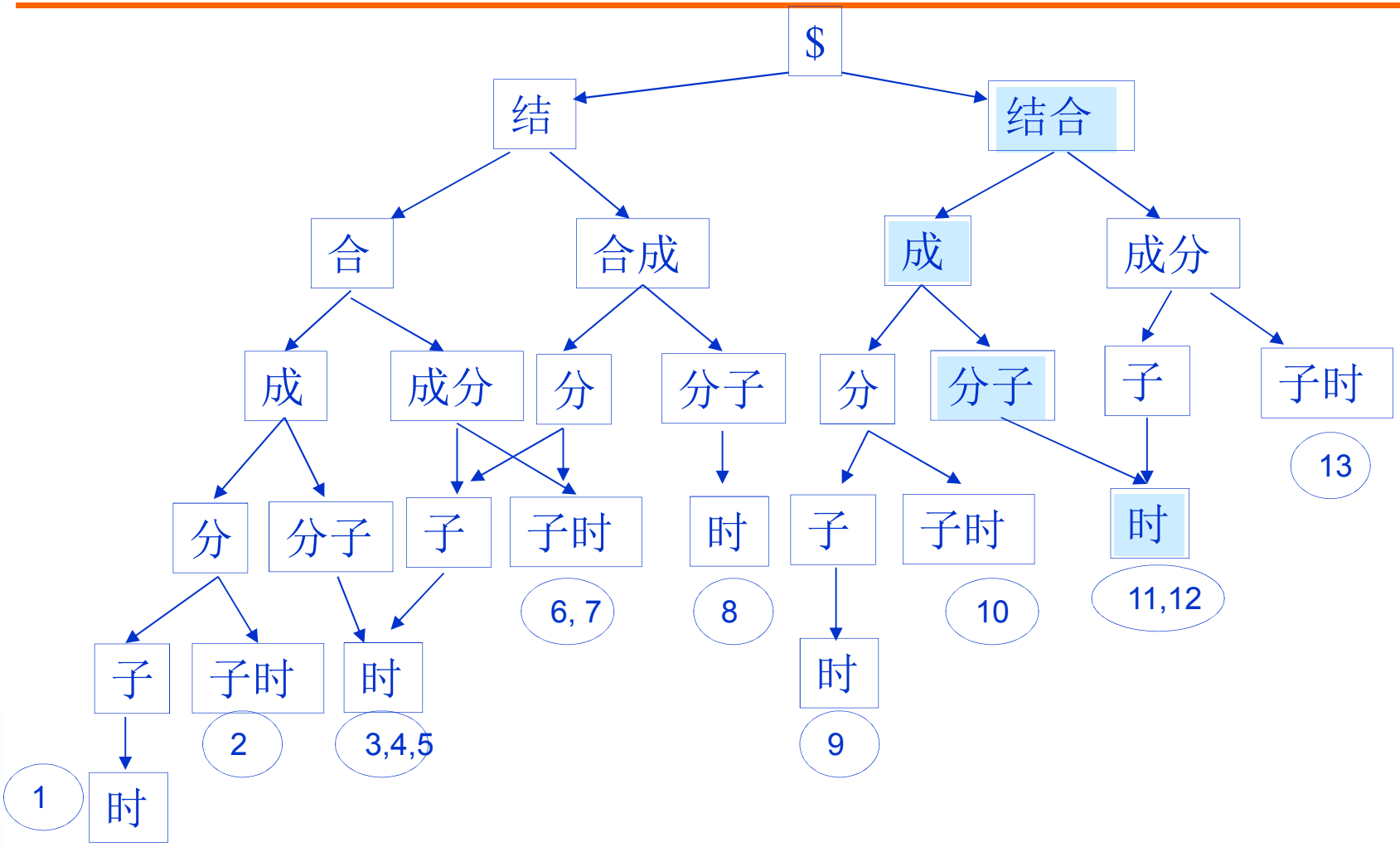
汉语切分的数据结构－词图



根据这个数据结构，可以把词法分析中的几种操作转化为：

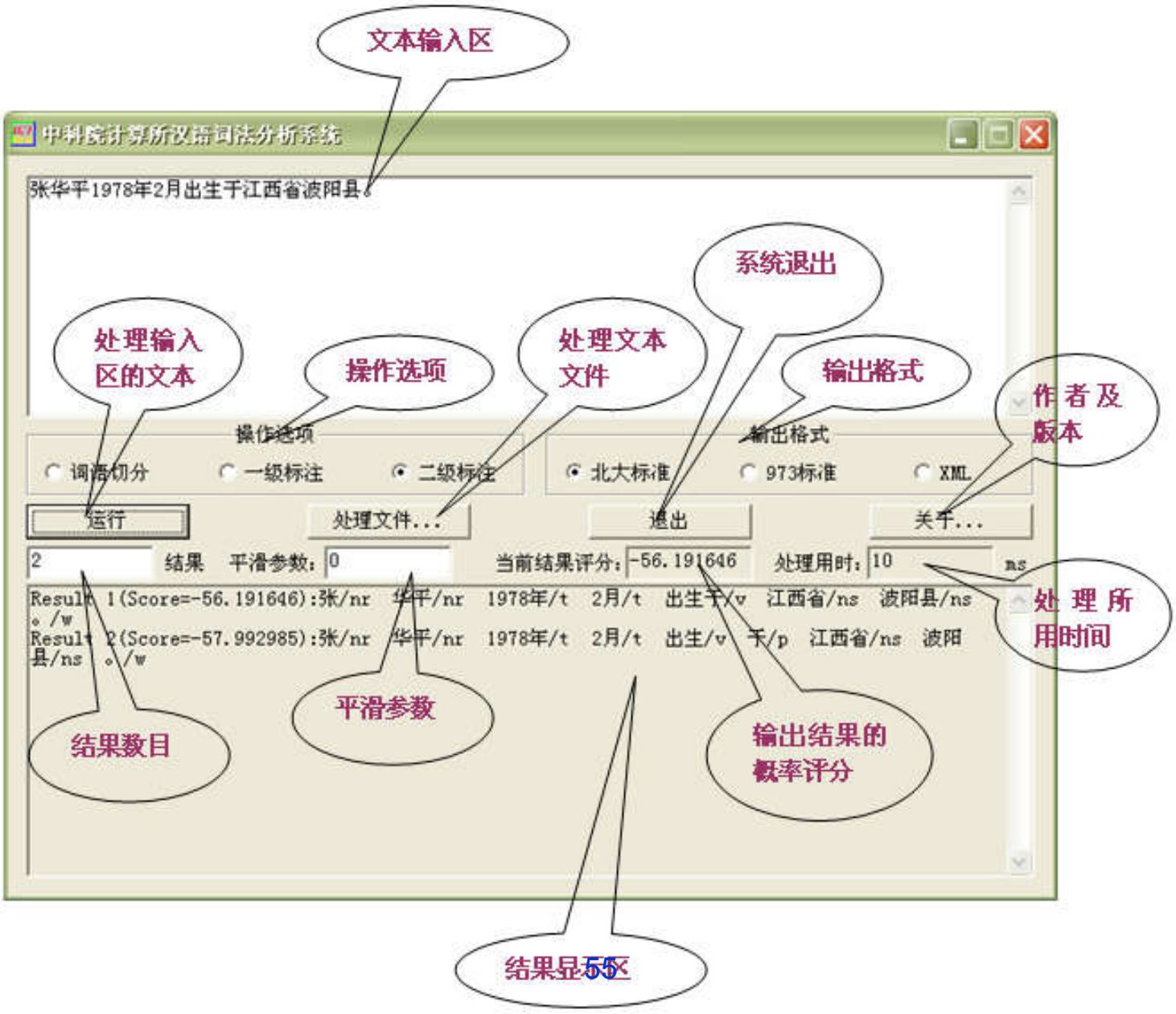
- 给词图上添加边（查词典，处理重叠词、离合词、前后缀和未定义词）
- 寻找一条起点S到终点E的最优路径（切分排歧）
- 给路径上的边加上标记（词性标注）

“结合成分子时”的全部分词路径示意



主要内容

- 基本概念
- 分词规范
- 分词词典
- 分词算法
- **分词系统**


文本输入区
中科院计算所汉语词法分析系统
张华平1978年2月出生于江西省波阳县。
系统退出
处理输入区的文本
操作选项
处理文本文件
输出格式
作者及版本
操作选项
输出格式
词语切分
一级标注
二级标注
北大标准
973标准
XML
运行
处理文件...
退出
关于...
2
结果
平滑参数:
0
当前结果评分:
-56.191646
处理用时:
10
ms
Result 1(Score=-56.191646):张/nr 华平/nr 1978年/t 2月/t 出生于/v 江西省/ns 波阳县/ns 。/w
Result 2(Score=-57.992985):张/nr 华平/nr 1978年/t 2月/t 出生/v 于/p 江西省/ns 波阳县/ns 。/w
处理所用时间
结果数目
平滑参数
输出结果的概率评分
结果显示区

常见分词系统

- ICTCLAS
  - 分词速度单机996KB/s
  - 分词精度98.45%
  - 采用C/C++编写，支持Linux、FreeBSD及Windows系列操作系统，支持C/C++、C#、Delphi、Java等主流的开发语言
  - 下载：
    - http://www.ictclas.org/

常见分词系统

- **Paoding**（庖丁解牛分词）
  - 基于Java的开源中文分词组件
  - 提供lucene和solr 接口
  - 在PIII 1G内存个人机器上，**1秒**可准确分词**100万**汉字
  - 采用基于 *不限制个数*的词典文件对文章进行有效切分，使能够将对词汇分类定义
  - 仅支持Java语言
  - http://code.google.com/p/paoding/

常见分词系统

- 斯坦福分词系统
  - 基于CRF（Condition Random Field，条件随机场）
  - 该系统提供了了两个分词数据模型
    - 宾州中文树库(Penn TreeBank)
    - 北京大学为sighan第二届中文分词竞赛提供的数据
  - 下载：
    - http://nlp.stanford.edu/software/segmenter.shtml

常见分词系统

- **SCWS**
  - 基于词频词典的机械中文分词引擎
  - 词频词典，并辅以一定的专有名称，人名，地名，数字年代等规则
  - 小范围测试:准确率在 90% ~ 95%
  - 切词时间大概是1.5MB文本/秒
  - 支持PHP4和PHP 5

海量分词

- 海量智能分词
- 海量信息技术有限公司开发
- 提供了自动分词和词性标注的功能。
- 其中“海量智能分词研究版”是一个免费的供学习、科研单位使用的版本
- 它没有提供源程序，但是提供了C++的调用接口，接口设计精良、调用方便，可以很方便地集成到自己需要的应用中。

分词困难

- 统计方法是主流
  - 但是，大量语言现象没有足够的统计信息
  - 齐普夫定律
    - 词数*词频=常量
      - 非常常用的词很少
      - 中频词的数量中等
      - 大量的低频词
    - 无法观测到足够数量的信息，很难预测那些在语料库中出现很少甚至不出现的行为

分词困难（续）

- 未登录词识别
  - 最大难点
- 标准
  - 有意义的基本单位是词
  - 但是词的定义不明确
    - 我们的 输入法中认为是输入单位
    - “牛肉”是词 “鳄鱼肉”、“龙肉”是不是词？

谢谢!