计算语言学

第6讲 词法分析 I

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院2002~2003学年第二学期课程讲义

语言的分类

传统语言学根据词的形态结构把语言分为三大类：

- 分析型语言：词基本上没有专门表示语法意义的附加成分，形态变化很少，语法关系靠词序和虚词来表示。如汉语、藏语等。
- 黏着型语言：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，一种语法意义也基本上由一个附加成分来表达，词根或词干跟附加成分的结合不紧密。如芬兰语、日语、蒙古语等。
- 屈折型语言：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，词根或词干跟词的附加成分结合得很紧密，往往不易截然分开。如：英语、德语和法语等。

屈折型语言的词法分析

- Tokenization：把字符串变成词串
  I'm a student. → I 'm a student .
- Lemmatization：对词的内部结构进行分析
  tokenization → token + ~ize + ~tion
  takes → take + ~s
  took → take + ~ed
- POS-Tagging: 词性标注

Tokenization

- 数字：123,456.78 90.7% 3/8 11/20/2000
- 缩略（包含不同的情况）：
  - 字母－点号－字母－点号组成的序列，比如：U.S. i.e. 等等；
  - 字母开头，最后以点号结束，比如：A. b. Mr. eds.prof. ；
- 包含非字母字符，比如：AT&T Micro$oft
- 带杠的词串，比如：three-years-old，one-third，so-called
- 带瞥号的词串，比如：I'm can't dog's let's
- 带空格的词串，比如："and so on"，"ad hoc"
- 其他：如网址（http://ict.ac.cn）、公式等

Tokenization问题

- 例外较多，跟文本来源有关
- 歧义现象（如点号的句子边界歧义）

数字的识别

数词的识别一般可以用有限状态自动机来实现

- 识别分数的正则表达式：
  - [0-9]+ / [0-9]
  - e.g. 12/21
- 识别百分数的正则表达式：
  - ([+ | -]) ? [0-9]+ (.) ? [0-9]* %
  - e.g. -5.9% 91%
- 识别十进制数字的正则表达式：
  - ([0-9]+(,)?)+ (. [0-9]+)?
  - e.g. 12,345

Tokenization算法

- 输入：一段文本
- 输出：单词串
- 算法：（略）

Lemmatization

屈折型语言的词语变化形式：

- 屈折变化：即由于单词在句子中所起的语法作用的不同而发生的词的形态变化，而单词的词性基本不变的现象，如（take，took，takes）。识别这种变化是词法分析的最基本的任务。
- 派生变化：即一个单词从另外一个不同类单词或词干衍生过来，如morphological ← morphology，英语中派生变化主要通过加前缀或后缀的形式构成；在其他语言中，如德语和俄语中，同时还伴有音的变化。
- 复合变化：两个或更多个单词以一定的方式组合成一个新的单词。这种变化形式比较灵活，如well-formed，6-year-old等等。

Lemmatization的目的：将上述变化还原

Lemmatization常见的问题

- 半规则变化
  - flied → fly + ~ed
  - rebelled → rebel + ~ed
- 不规则变化
  - good, better, best
  - child, children
- 歧义现象
  - better → good + ~er or well + ~er ?
  - works → work + ~s or works ?

Lemmatization规则示例 1

- 名词复数

```
*s  → *, (PLUR)
*es → *, (PLUR)
*ies → *y, (PLUR)
```

- 动词第三人称单数

```
*s → *    (SINGULAR) (THIRDPERSON)
*es → *   (SINGULAR) (THIRDPERSON)
*ies → *y (SINGULAR) (THIRDPERSON)
```

Lemmatization规则示例 2

- 动词现在分词
  - *ing → * (VING)
  - *ing → *e (VING)
  - *ying→ *ie (VING)
  - ??ing →? (VING)
- 动词过去分词、过去式
  - *ed → * (PAST,VEN)
  - *ed → *e (PAST,VEN)
  - *ied → *y (PAST,VEN)
  - *??ed → *? (PAST,VEN)

Lemmatization算法

- 输入：一个单词
- 输出：一个或多个单词，其中每个单词还原为原形加前后缀（可以有多个）
- 算法：（略）

Lemmatization要做到何种程度

- 词干层。如：
  impossibilities→impossibility+ies
- 词根层。如：
  impossibilities→im+poss+ibil+it+ies
- 分析程度取决于自然语言处理系统的深度：
  - 不解决未定义词，分析到词干层
  - 解决未定义词，要分析到词根层。

汉语词法分析所面临的问题

- 重叠词、离合词、词缀
- 汉语词语的切分歧义
- 汉语未定义词
- 词性标注

汉语双字形容词的重叠形式

形容词(AB)	ABAB式	AABB式	A里AB式
高兴	高兴高兴	高高兴兴	
明白	明白明白	明明白白	
热闹	热闹热闹	热热闹闹	
潇洒	潇洒潇洒	潇潇洒洒	
糊涂		糊糊涂涂	糊里糊涂
流气			流里流气
粘乎	粘乎粘乎	粘粘乎乎	
凉快	凉快凉快	凉凉快快	

汉语单字形容词的重叠形式

形容词（A）	AA式	ABB式	ABCD式
黑	黑黑	黑压压	黑不溜秋
白	白白	白花花	白不呲咧
红	红红	红彤彤	
亮	亮亮	亮晶晶	
恶		恶狠狠	
香	香香	香喷喷	
滑	滑滑	滑溜溜	

汉语双字动词的重叠形式

动词(AB)	ABAB式	AABB式
研究	研究研究	
讨论	讨论讨论	
哆嗦		哆哆嗦嗦
唠叨		唠唠叨叨
嘀咕		嘀嘀咕咕

汉语单字动词的重叠形式

动词（V）	VV式	V一V式	V了V式	V了一V式
听	听听	听一听	听了听	听了一听
想	想想	想一想	想了想	想了一想
玩	玩玩	玩一玩	玩了玩	玩了一玩
醒	醒醒	醒一醒		
试	试试	试一试	试了试	试了一试
笑	笑笑	笑一笑	笑了笑	笑了一笑
讲	讲讲	讲一讲	讲了讲	讲了一讲

汉语其他词类的重叠形式

- 名词
  - 哥哥，人人
  - 山山水水，是是非非，方方面面，头头脑脑
- 数词
  - 一一做了回答，两两结伴而来
- 量词
  - 个个都是好样的，回回考满分
- 副词
  - 常常，仅仅，的的确确

汉语重叠词的特点

- 汉语词能否重叠具有很强的个性特点
  - 研究研究√
  - 工作工作×
- 有些词重叠后词性发生了变化
  - 形容词重叠后一般成为状态词
  - 个别量词重叠后可以成为其他词性
    - 回回：副词
    - 个个：名词

汉语词缀

- 前缀
  - 老鹰、老虎、老三、老王
  - 超豪华、超标准、超高速
  - 非党员
- 后缀
  - 骨头、砖头、甜头、苦头、盼头、想头
  - 桌子、椅子、孩子、票子、房子
  - 文学家、指挥家、艺术家
  - 科学性、可能性、学术性
  - 碗儿、花儿、玩儿、份儿、片儿

汉语离合词

- 汉语动词存在离合词现象
  - 游泳：游了一会儿泳
  - 理发：发理了没有
  - 担心：担什么心
  - 洗澡：洗了个热水澡
- 白硕的解释：语义重心偏移
  - 动词虚化（类似英语DO）
  - 语义重心落在后面的名词性语素上
  - 游泳：游了一会儿泳：DO了一会儿游泳
  - 理发：发理了没有：理发DO了没有
  - 担心：担什么心：DO什么担心
  - 洗澡：洗了个热水澡：DO了一个热水洗澡

处理识别词形变化的规则 1

@@ VV -- v [重叠形式:VV] << V -- v
@@ UVUV -- v [重叠形式:UVUV] << UV -- v
@@ V了V -- v [重叠形式:V了V] << V -- v
@@ V一V -- v [重叠形式:V一V] << V -- v
@@ V了N -- v [重叠形式:V了N] << VN -- v [趋向动词:否]
@@ VVN -- v [重叠形式:VVN] << VN -- v
@@ V过N -- v [重叠形式:V过N] << VN -- v [趋向动词:否]
@@ V了一N -- v [重叠形式:V了一N] << VN -- v
@@ V过一N -- v [重叠形式:V过一N] << VN -- v
@@ V不了N -- v [重叠形式:V不了N] << VN -- v

处理识别词形变化的规则2

@@ AA -- a [重叠形式:AA] << A -- a
@@ AABB -- a [重叠形式:AABB] << AB -- a
@@ ABAB -- a [重叠形式:ABAB] << AB -- a
@@ DD -- d [重叠形式:DD] << D -- d
@@ R俩 -- r << R们 -- r
@@ N儿 -- n [重叠形式:N儿] << N -- n
@@ MN儿 -- n [重叠形式:MN儿] << MN -- n

汉语的切分歧义

- 交集型歧义（交叉型歧义）：如果字串abc既可切分为ab/c，又可切分为a/bc。其中a，ab，c和bc是词
  - 有意见：我 对 他 有 意见。 总统 有意 见 他。
- 组合型歧义（覆盖型歧义）：若ab为词，而a和b在句子中又可分别单独成词
  - 马上：我 马上 就 来。 他 从 马 上 下来。
  - 将来：我 将来 要 上 大学。 我 将 来 上海。
- 混合型歧义：由交集型歧义和组合型歧义自身嵌套或两者交叉组合而产生的歧义
  - 人才能：这样 的 人才 能 经受 住 考验。
  - 人才能：这样 的 人 才能 经受 住 考验。
  - 人才能：这样 的 人 才 能 经受 住 考验。

交集型歧义字段的链长

- 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
  - 链长为2：结合成分
  - 链长为3：为人民工作
  - 链长为4：中国产品质量　结合成分子时
  - 链长为6：努力学习语法规则
  - 链长为7：治理解放大道路面积水

真歧义和伪歧义

- 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”
- 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”

真实语料中歧义字段的分布

刘开瑛，2000，《中文文本自动分词和标注》，商务印书馆，第65页。

（500万新闻语料的统计结果）

链长	1	2	3	4	5	6	7	8	总计
词次数	47402	28790	1217	608	29	19	2	1	78248
比例	50.58	47.02	1.56	0.78	0.04	0.02	0.00	0.00	100
字段数	12686	10131	743	324	22	5	2	1	23914
比例	53.05	42.36	3.11	1.35	0.09	0.02	0.01	0.01	100

词语切分标准 1

- 建立汉语词语切分标准的必要性
  - 汉语词语定义不明确
    - 牛肉是词，鸡肉是不是？
    - 打倒是词，打死、打伤、饿死、涂黑是不是？
  - 为操作的方便，必须确定统一的标准或规范
  - 采用“分词单位”的说法
- 问题
  - 取舍理由不够充分，人为色彩过重
  - 过于复杂，难于把握

词语切分标准 2

- 相关的标准
  - 《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
  - 《人民日报》语料库词语切分规范
  - ……

词语切分标准 3（例）

《人民日报》标注语料库词语切分规范（述补结构的切分）

未收入词典的双音节述补结构，若拆开各是一个词，通常作为两个切分单位。

走/v 到/v，撞/v 上/v， 调/v 好/a，坐/v 稳/a

若拆开了，其中至少有一个是语素，通常就不切分，作为一个切分单位。

形成/v， 鼓动/v， 说明/v， 震动/v

双音节的述补结构中间插入“得”或“不”一般应予切分，

走/v 得/u 到/v，走/v 不/d 到/v，安/v 得/u 上/v，安/v 不/d 上/v

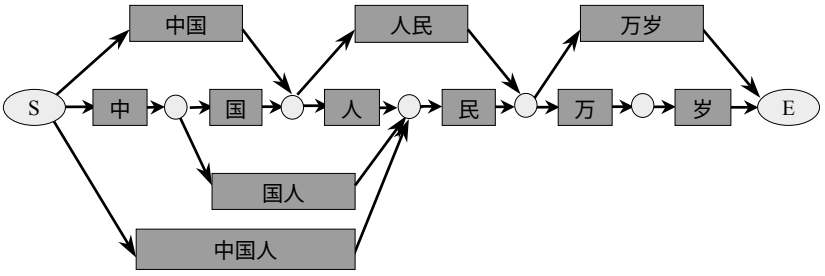
但是如果去掉“得”或“不”后，前后两个字不构成一个词的，则作为一个分词单位。

来得及/v，来不及/v，对得起/v，对不起/v，说得过去/l，说不过去/l

有的去掉“得”或“不”后虽然是一个合成词，但其中至少有一个是语素，拆开了却是难以理解的，仍作为一个切分单位

形得成/v，形不成/v

汉语切分的数据结构－词图



根据这个数据结构，我们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）；

汉语切分算法

- 基于词典的机械切分算法
  - 全切分
  - 最大匹配方法
  - 最短路径方法
  - 交叉歧义检测法
  - 基于记忆的交叉歧义排除法
- 基于规则的切分算法
- 基于统计的切分算法
  - N元语法
  - 最大压缩方法

全切分方法

- 给出所有的切分结果
- 算法（略）
- 算法的时间复杂度随着句子长度的增加呈指数增长

最大匹配方法 1

- 正向最大匹配（MM）
  - 自左往右
  - 每次取最长词
- 逆向最大匹配（RMM)
  - 自右往左
  - 每次取最长词
- 双向最大匹配
  - 依次采用正向和逆向最大匹配
  - 如果结果一致则输出
  - 如果结果不一致再用其他方法排歧

最大匹配方法 2

- 优点
  - 简单、快速
  - 在某些应用场合已经足够
- 缺点
  - 单向最大匹配会忽略交集型歧义和组合型歧义
    幼儿园 地 节目 ／ 独立自主 和平 等 互利 的 原则
  - 双向最大匹配会忽略链长为偶数的交集型歧义和组合型歧义
    原子 结合 成分 子时 ／ 他 从 马上 下来

最短路径方法

- 基本思想：
  - 在词图上选择一条词数最少的路径
- 算法：
  - 动态规划算法
- 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)
  - 最短路径：独立自主 和 平等互利 的 原则(5)
- 缺点：忽略了所有覆盖歧义，也无法解决大部分交叉歧义
  - 结合 成分 子时

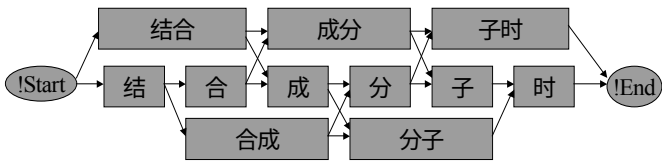
交叉歧义检测 1

- 王显芳，杜利民，一种可以检测所有交叉歧义的汉语切分算法，电子学报，2003（已录用）
- 可以输出一个句子所有可能的交叉歧义切分结果，但忽略所有覆盖歧义切分结果
- 算法（略）
- 优点：
  - 高效时间复杂度：线性
  - 可以明确将一个句子的交叉歧义和覆盖歧义分开，以便在后续步骤中采用不同策略进行处理

交叉歧义检测 2

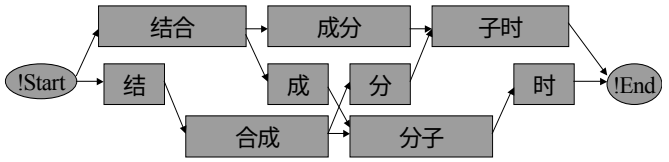
全切分：

13种结果

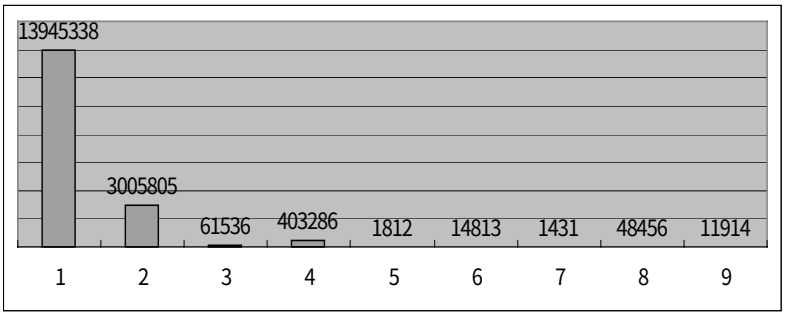


交叉歧义：

4种结果



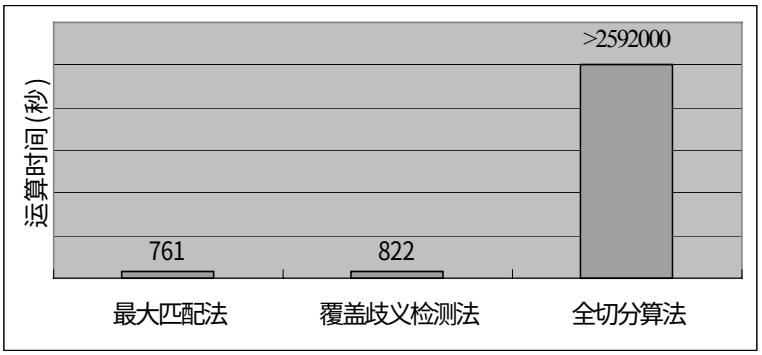
交叉歧义检测 3



最大无覆盖歧义切分方式个数为1到8和大于8的句子个数

根据十一年《人民日报》等520M语料

交叉歧义检测 4



最大匹配法、覆盖歧义检测法和全切分算法所需要的时间

基于记忆的交叉歧义排除法

- 孙茂松，左正平，邹嘉彦，1999, 高频最大交集型歧义切分字段在汉语自动分词中的作用，中文信息学报，Vol.13，No.1，1999
- 该文考察了一亿字的语料，发现交集型歧义字段的分布非常集中。其中在总共的22万多个交集型歧义字段中，高频的4,619个交集型歧义字段占所有歧义切分字段的59.20%。而这些高频歧义切分字段中，又有4,279个字段是伪歧义字段，也就是说，实际的语料中只可能出现一种切分结果。这样，仅仅通过基于记忆的方法，保存一种伪歧义切分字段表，就可以使交集型歧义切分的正确率达到53%，再加上那些有严重偏向性的真歧义字段，交集型歧义切分的正确率可以达到58.58%。

基于规则的切分方法

- @@ 高峰(A+B, AB)
  CONDITION FIND(L,NEXT,X){%X.yx=最|更} SELECT 1
  OTHERWISE SELECT 2
- @@ 分成(A+B, AB)
  CONDITION FIND(R,NEXT,X){%X.ccat=m} SELECT 1
  OTHERWISE SELECT 3
- @@ 是因为(A+B,AB)
  CONDITION FIND(L,FAR,X){%X.yx=之所以} SELECT 2
  OTHERWISE SELECT 1
- @@ *(A+BC, AB+C)
  CONDITION %A.ccat =p, %BC.ccat =n |f |s |t SELECT 1
- @@ *(A+BC, AB+C)
  CONDITION %A.ccat=v, %BC.ccat=b|d|t, %AB.ccat=r, %C.ccat=n
  SELECT 2

基于N元语法的切分排歧 1

$$
\begin{aligned}
W^* &= \arg\max_W P(W \mid C) \\
&= \arg\max_W \frac{P(C \mid W)P(W)}{P(C)} \\
&= \arg\max_W P(W) \\
&= \arg\max_W p(w_1 \ldots w_{N-1}) \prod_{i=N}^{l} p(w_i \mid w_{i-N+1} \ldots w_{i-1})
\end{aligned}
$$

算法：动态规划（类似于HMM解码问题的Viterbi算法）

基于N元语法的切分排歧 2

- 采用一元语法
  - 即把切分路径上每一个词的词频相乘得到该切分路径的概率
  - 把词频的负对数理解成“代价”，这种方法也可以理解为最短路径法的一种扩充
  - 正确率可达到92%
  - 简便易行，效果一般好于基于词表的方法

基于N元语法的切分排歧 3

- 采用三元语法
  - 王显芳，2001，利用覆盖歧义检测法和统计语言模型进行汉语自动分词
  - 高山，张艳，徐波，宗成庆，韩兆兵，张仰森，2001，基于三元统计模型的汉语分词标注一体化研究，全国第五届计算语言学联合学术会议（JSCL2001）
  - 在不考虑未定义词的情况下，就可以将切分的正确率提高到98%以上。

基于最大压缩的切分算法

- W.J.Teahan,Yingying Wen,Rodger McNab,Ian H. Witten, A compression-based algorithm for Chinese word segmentation, *computational linguistics*, 26(3):375-393, 2000

分词系统的评价

- 方法一：严格按照某种规范进行评价
  - 算法简单
  - 不够合理
- 方法二：具有一定的容忍度
  - “鸡肉”切成一个词或两个词都算正确
  - 较为合理
  - 算法与数据结构都较为复杂，或者引入人工
  - 测试数据的构造比较费时

未定义词的类型

- 中国人名：李素丽 老张 李四 王二麻子
- 中国地名：定福庄 白沟 三义庙 韩村河 马甸
- 翻译人名：乔治·布什 叶利钦 包法利夫人 酒井法子
- 翻译地名：阿尔卑斯山 新奥尔良 约克郡
- 机构名：方正公司 联想集团 国际卫生组织 外贸部
- 商标字号：非常可乐 乐凯 波导 杉杉 同仁堂
- 专业术语：万维网 主机板 模态逻辑 贝叶斯算法
- 缩略语：三个代表 五讲四美 打假 扫黄打非 计生办
- 新词语：卡拉OK 波波族 美刀 港刀

未定义词识别的困难

- 未定义词没有明确边界
- 未定义词的构成单元（汉字）本身都可以独立成词

未定义词识别的一般方法

- 每一类未定义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）
  - 重复出现规律

未定义词识别的研究进展

- 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

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%。

中国人名的内部构成规律 3

- 中国人名各组成部分的组合规律
  - 姓＋名
  - 姓
  - 名
  - 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：女士、教授、经理、小姐、总理
- 地名或机构名：
  - 前：静海县大丘庄禹作敏
- 的字结构
  - 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]
- 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
- 人名地名冲突
  - 河北省*刘庄*

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

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”。

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

机构名的内部构成规律 1

- 机构名一般都是定中结构
- 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

未定义词识别的一般方法

- 各种不同类型的未定义词识别方法思想大同小异，但实现时各有侧重
- 各种不同类型的未定义词识别都需要收集大量数据，建立不同的数据模型
- 常用的方法包括
  - 规则方法：人工总结或归纳出一些判别规则，并用程序实现
  - 统计方法：建立统计模型，通过人工标注语料库进行参数训练

未定义词识别的评价标准

- 正确率 $\mathrm{Pr}\,ecision = \dfrac{\text{正确识别的未定义词数}}{\text{所有识别的未定义词数}}$
- 召回率 $\mathrm{Re}\,call = \dfrac{\text{正确识别的未定义词数}}{\text{需要识别的未定义词数}}$
- F-Score
  一般取β＝1 $F - Score = \dfrac{(\beta^2+1)\times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\beta^2 \times \mathrm{Pr}\,ecision + \mathrm{Re}\,call}$

由于正确率和召回率一般存在反比关系，
因此常用F-Score来作总体评价

基于角色标注的人名识别 1

Huaping Zhang, Qun Liu, Hao Zhang,Xueqi Cheng, Automatic Recognition of Chinese Unknown Words Based on Role Tagging, SigHan Workshop, attached with 19th International Conference on Computational Linguistics, Taipei, 2002.8.

基于角色标注的人名识别 2

中国人名识别的角色定义

角色	意义	例子
B	姓氏	*张*华平先生
C	双名的首字	张*华*平先生
D	双名的末字	张华*平*先生
E	单名	张*浩*说：“我是一个好人”
F	前缀	*老*刘、*小*李
G	后缀	王*总*、刘*老*、肖*氏*、吴*妈*、叶*帅*
K	人名的上文	又*来到*于洪洋的家。
L	人名的下文	新华社记者黄文*煊*
M	两个中国人名之间的成分	编剧邵钧林*和*稽道青说
U	人名的上文和姓成词	这里*有关*天培的壮烈
V	人名的末字和下文成词	龚学*平等*领导，邓颖*超生*前
X	姓与双名的首字成词	*王国*维、
Y	姓与单名成词	*高峰*、*汪洋*
Z	双名本身成词	张*朝阳*
A	以上之外其他的角色	

基于角色标注的人名识别 3

- 例子
  - 人名识别前的切分结果：
    馆/内/陈列/周/恩/来/和/邓/颖/超生/前/使用/过/的/物品/。
  - 角色标注后的结果：
    馆/A 内/A 陈列/K 周/B 恩/C 来/D 和/M 邓/B 颖/C 超生/V 前/A 使用/A 过/A 的/A 物品/A 。/A

基于角色标注的人名识别 4

- 采用隐马尔科夫模型（HMM）进行角色标注
  - 将“角色”理解为HMM中的“状态”
  - 将人名识别前切分出的词理解为HMM中的“观察值”
  - 将已有的词标注语料库（《人民日报》语料库）转换为角色标注语料库
  - 利用角色标注语料库对HMM模型参数进行训练

基于角色标注的人名识别 5

- 语料库的转换
  - 《人民日报》语料库原始姓氏
    政务司/n 司长/n 陈/nr 方/nr 安生/nr 出任/v
    委员会/n 主席/n
  - 转换后的形式
    政务司/A 司长/K 陈/B 方/B 安/C 生/D 出任/L
    委员会/A 主席/A
  - 统计得到状态转移矩阵和输出矩阵

基于角色标注的人名识别 6

- 角色分裂
  - 在人名识别之前，我们要对角色U和V进行分裂处理。相应地分裂为KB、DL或者EL
  - 例子：
    - 馆/内/陈列/周/恩/来/和/邓/颖/超生/前/使用/过/的/物品/。
    - 分裂前：AAKBCDMBC*V*AAAAAA
    - 分裂后：AAKBCDMBC*DL*AAAAAA
- 模式匹配：得到人名
  - BBCD, BBE, BBZ, BCD,
    BEE,BE,BG,BXD,BZ,CD,EE,FB, Y,XD
  - 例子： 模式BCD：周恩来，邓颖超

基于角色标注的人名识别 7

类别	封闭测试语料1	封闭测试语料2	开放测试语料
来源：《人民日报》	98年1月	98年2月1日-20日	98年2月20日-28日
语料库大小（字节）	8,621K	6,185K	2,605K
实际人名数	13360	7224	2967
识别出的人名数	17505	10929	4259
正确识别的人名数	13079	7106	2739
准确率	74.72%	65.02	64.32%
召回率	97.90%	98.37%	92.32%
F值	84.75%	78.29%	75.81%

测试结果

基于角色标注的人名识别 8

- 结果分析
  - 完全真实的测试环境：没有剔除不含人名的句子
  - 测试结果仅针对词典中未定义的人名（如果考虑词典中已有的人名，正确率和召回率都将达到95%以上）
  - 实验规模大
  - 还有改进的余地
  - 基于角色标注的方法可用于各种未定义词识别

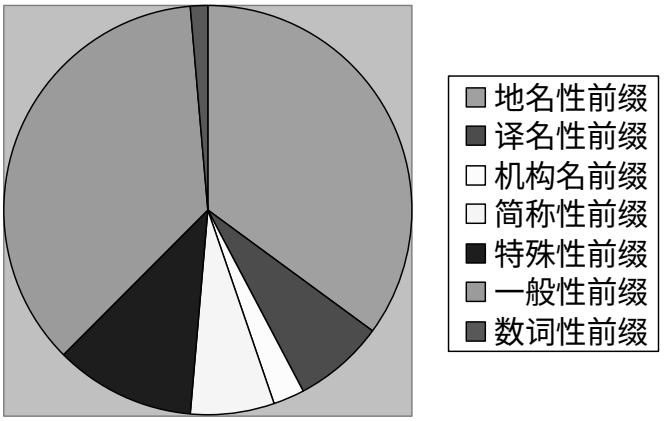
基于角色标注的音译名识别

代码	含义	例子
B	首字	*克*/林/顿
C	中字	史/*蒂*/*芬*·/*斯*/*皮*/*尔*/*伯*/格
D	尾字	克/林/*顿*
K	左邻（上文）	
L	右邻（下文）	
M	两个音译名之间的连接	
A	其他	
……		

基于角色标注的机构名识别 1

中文机构名称构成角色表		
角色	意义	例子
A	上文	*参与*亚太经合组织的活动
B	下文	中央电视台*报道*
X	连接词	北京电视台*和*天津电视台
C	特征词的一般性前缀	
F	特征词的译名性前缀	美国*摩托罗拉*公司
G	特征词的地名性前缀	交通银行*北京*分行
H	特征词的机构名前缀	*中共中央*顾问委员会
I	特征词的特殊性前缀	*中央*电视台
J	特征词的简称性前缀	
D	机构名的特征词	
Z	非机构名成份	

基于角色标注的机构名识别 2



基于角色标注的机构名识别 3

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报1月	7836	8476	6317	74.5	80.6	77.5
人民日报6月	9065	10216	7136	69.9	78.7	74.0

- 训练语料都是《人民日报》1998年1～5月语料
- 实验结果分析
  - 角色集合的选取对识别的结果至关重要，要反复尝试
  - 加入机构名识别后人名地名的识别正确率都有所提高

复习思考题

- 请判断下面的句子属于哪一种歧义现象，如果是交集型切分歧义，请给出歧义字段的链长：
  - 难过----我们家门前的大水沟很难过。
  - 如果----罐头不如果汁营养丰富。
  - 天真----今天真热，是游泳的好日子。
  - 十分----妹妹的数学只考十分，真丢脸。
  - 从容----我做事情，都是先从容易的做起。
  - 人参----老师说明天每个人参加大队接力时，一定要尽力。
- 请为汉语文本中经常出现的数字和日期的识别分别构造一部有限状态自动机
- 有人提出，改变汉语书写习惯，采用词儿连写，你认为这种做法有什么优缺点？

复习思考题

- 设计一种基于转换的错误驱动的汉语切分算法
- 请给出基于N元语法的动态规划分词方法的算法描述