



中文自动分词

计算机科学与技术学院

刘秉权

新技术楼612室，电话：86413322

Email: liubq@insun.hit.edu.cn



主要内容

- 分词的提出
- 分词歧义
- 分词规范
- 主要分词方法
- 生词识别



分词的提出和定义

- 汉语文本是基于单字的，汉语的书面表达方式也是以汉字作为最小单位的，词与词之间没有显性的界限标志，因此分词是汉语文本分析处理中首先要解决的问题
- 添加合适的显性的词语边界标志使得所形成的词串反映句子的本意，这个过程就是通常所说的分词（Chinese Word Segmentation）



分词的意义

- 正确的机器自动分词是正确的中文信息处理的基础
 - 文本检索
 - 和服 | 务 | 于三日后裁制完毕，并呈送将军府中。
 - 王府饭店的设施 | 和 | 服务 | 是一流的。
如果不分词或者“和服务”分词有误，都会导致荒谬的检索结果。
 - 文语转换
 - 他们是来 | 查 | 金泰 | 撞人那件事的。（“查”读音为cha）
 - 行侠仗义的 | 查金泰 | 远近闻名。（“查”读音为zha）



分词面临的主要难题

- 如何面向大规模开放应用是汉语分词研究当前面临的主要问题
 - 词语边界歧义处理
 - 如何识别未登录词
 - 如何低廉地获取语言学知识
 - 实时性应用中的效率问题



分词歧义

- 交集型切分歧义
- 组合型切分歧义



交集型切分歧义

- 汉字串AJB被称作交集型切分歧义，如果满足AJ、JB同时为词(A、J、B分别为汉字串)。此时汉字串J被称作交集串。
 - [例] “**美国会**通过对台售武法案”
 - [例] “**乒乓球**拍卖完了”
 - [例] “**结合**成分子”
 - 结合|成分|子|
 - 结合|成|分子|
 - 结|合成|分子|



组合型切分歧义

- 汉字串AB被称作组合型切分歧义，如果满足条件：A、B、AB同时为词
 - [例]组合型切分歧义：“起身”
 - 他站 | 起 | 身 | 来。
 - 他明天 | 起身 | 去北京。



更多组合歧义实例

- "把手","这个门的把手坏了好几天了","你把手抬高一点儿"
- "本书","本书讨论的问题是一个老生常谈的问题","那本书写得非常精彩"
- "并排","这条马路可以并排行驶四辆大卡车","教务科指定了专任讲师并排好了课程时间表"
- "病痛","人身上哪怕有一小病痛,都会影响到工作学习","这种病痛起来真要人命"
- "病因","这种病的病因到目前为止医学界都不清楚","她的病因我而起,就由我来解决吧"
- "不大","他平时不大抽烟,想不到也得了肺癌","他年纪不大鬼点子却不少"
- "不过","本来他是想去赴宴的,不过这两天胃口不好,就只得做罢了","这次考试要还是不过,我就自杀"
- "不要","在公众场合不要发出这种声音","我不要回我的东西誓不罢休"
- "才能","没有出众的才能就无法在竞争中站稳脚跟","掌握新技术才能立于不败之地"
- "炒菜","他进馆子就要了两个炒菜一瓶二锅头","小王炒菜的手艺不错"
- "穿着","我一看他的穿着打扮就知道他不是等闲之辈","她今天是穿着一身礼服出去的"
- "词组","要练好地道的英语口语,必须熟练掌握一些常用词组的用法","使用语言的过程基本上就是选词组句的过程"
- "打包","我们的行李都已经打包了,很难找到那份文件","我有一打包好的丝袜"
- "打手","这个大款养了一群打手替他开路","小明妈妈打小明的時候不仅打手还打屁股"
- "大路","你顺着这条大路一直往前走就到了","长这么大路都不会走的孩子我还是第一回见到"
- "跟头","他昨天一连摔了两个跟头,都摔得不轻","那位老大爷的脚跟头都摔破了"
- "鬼才","这位是有“文坛鬼才”之称的魏先生","鬼才相信他说的话呢"
- "过奖","您过奖了,我做的只是我应该做的事情","他参加过奥林匹克数学竞赛,还得过奖呢"



“真歧义”和“伪歧义”

- 真歧义指存在两种或两种以上的可实现的切分形式，如句子“必须/加强/企业/中/国有/资产/的/管理/”和“中国/有/能力/解决/香港/问题/”中的字段“中国有”是一种真歧义
- 伪歧义一般只有一种正确的切分形式，如“建设/有”、“中国/人民”、“各/地方”、“本/地区”等



未登录词

- 虽然一般的词典都能覆盖大多数的词语，但有相当一部分的词语不可能穷尽地收入系统词典中，这些词语称为未登录词或新词
- 分类：
 - 专有名词：中文人名、地名、机构名称、外国译名、时间词
 - 重叠词：“高高兴兴”、“研究研究”
 - 派生词：“一次性用品”
 - 与领域相关的术语：“云计算”、“转基因”
 - 网络热词：“神马”、“浮云”、“有木有”、“四袋苹果”、“宅男”、“宅女”、“低碳哥”



分词规范

- 词是自然语言的一种客观存在
- 汉语书写过程中并不分词连写，对词组和词、单字语素和单字词的划分因人而异，甚至因时而异
- 汉语信息处理需要制订统一的分词标准，否则将严重影响计算机的处理
- 《信息处理用现代汉语分词规范及自动分词方法》：结合紧密、使用频繁



具体的分词标准实例

- 1 二字或三字词，以及结合紧密、使用稳定的：
发展 可爱 红旗 对不起 自行车 青霉素

- 2 四字成语一律为分词单位：胸有成竹 欣欣向荣

四字词或结合紧密、使用稳定的四字词组：社会主义 春夏秋冬 由此可见

- 3 五字和五字以上的谚语、格言等，分开后如不违背原有组合的意义，应予切分：

时间/就/是/生命/

失败/是/成功/之/母



具体的分词标准实例

- 4 结合紧密、使用稳定的词组则不予切分:不管三七二十一
- 5 惯用语和有转义的词或词组, 在转义的语言环境下, 一律为分词单位:
 妇女能顶/半边天/
 他真小气, 象个/铁公鸡/
- 6 略语一律为分词单位:科技 奥运会 工农业
- 7 分词单位加形成儿化音的“儿”:花儿 悄悄儿 玩儿



具体的分词标准实例

- 8 阿拉伯数字等，仍保留原有形式:1234
7890
- 9 现代汉语中其它语言的汉字音译外来词，不予切分:巧克力 吉普
- 10 不同的语言环境中的同形异构现象，按照具体语言环境的语义进行切分:
把/手/抬起来
这个/把手/是木制的



常见的动词分词规范

- 1 动词前的否定副词一律单独切分:不/写 不/能 没/研究 未/完成
- 2 用肯定加否定的形式表示疑问的动词词组一律切分, 不完整的则不予切分:说/没/说 看/不/看 相信/不/相信
- 3 动宾结构的词或结合紧密、使用稳定的:开会 跳舞 解决/吃饭/问题 孩子该/念书/了
- 4 结合不紧密或有众多与之相同结构词组的动宾词组一律切分:吃/鱼 学/滑冰 写/信



常见的动词分词规范

- 5 动宾结构的词或词组如中间插入其它成分，则应予切分:吃/两/顿/饭 跳/新疆/舞
- 6 动补结构的二字词或结合紧密、使用稳定的二字动补词组，不予切分:打倒 提高 加长 做好
- 7 “2+1,1”或“1+2”结构的动补词组一律切分:整理/好 说/清楚 解释/清楚 打/得/倒 提/不/高
- 8 偏正结构的词，以及结合紧密的词不予切分:胡闹 瞎说 死记



常见的动词分词规范

- **9** 复合趋向动词一律为分词单位:出去
进来
当插入“得、不”时应予切分:出/得/去
进/不/来
- **10** 动词与趋向动词结合的词组一律切分:
寄/来 跑/出去
- **11** 多字动词无连词并列, 一律切分:调查
/研究 宣传/鼓动



■ 问题：如何分词？



主要的分词方法

- 简单的模式匹配：正向最大匹配、逆向最大匹配法、双向匹配法
- 基于规则的方法：最少分词算法
- 基于统计的方法：统计语言模型分词、串频统计和词形匹配相结合的汉语自动分词、无词典分词



正向最大匹配分词(Forward Maximum Matching method, FMM)

■ 基本算法:

- 1.设自动分词词典中最长词条所含汉字个数为I;
- 2.取被处理材料当前字符串序数中的I个字作为匹配字段，查找分词词典。若词典中有这样的I字词，则匹配成功，匹配字段作为一个词被切分出来，转6;
- 3.如果词典中找不到这样的I字词，则匹配失败;
- 4.匹配字段去掉最后一个汉字，I--;
- 5.重复2-4，直至切分成功为止;
- 6.I重新赋初值，转2，直到切分出所有词为止。



实现细节

- 词库的组织形式
- 搜索算法



分析

- “市场/中国/有/企业/才能/发展/”
- 对交叉歧义和组合歧义没有什么好的解决办法
- 错误切分率为1 / 169
- 往往不单独使用，而是与其它方法配合使用



逆向最大匹配分词(Backward Maximum Matching method, BMM法)

- 分词过程与FMM方法相同，不过是从句子(或文章)末尾开始处理，每次匹配不成功时去掉的是前面的一个汉字
- “市场/中/国有/企业/才能/发展/
- 实验表明：逆向最大匹配法比正向最大匹配法更有效，错误切分率为1 / 245

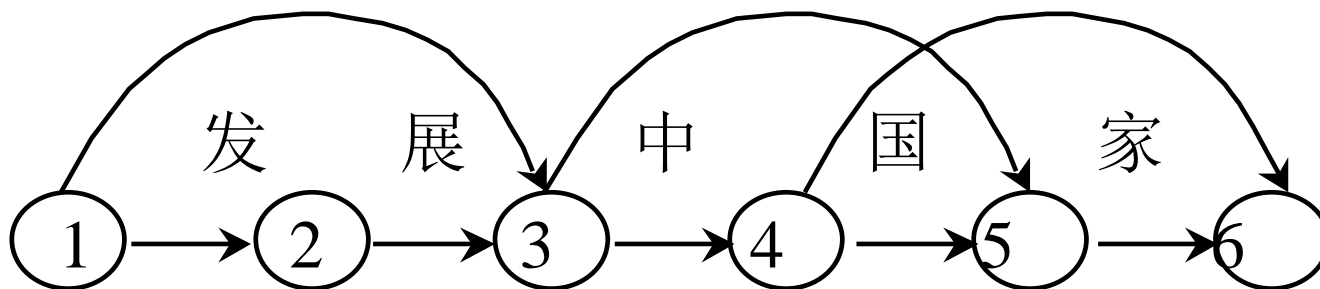


双向匹配法（Bi-direction Matching method, BM法）

- 比较FMM法与BMM法的切分结果，从而决定正确的切分
- 可以识别出分词中的交叉歧义

最少分词问题

- 分词结果中含词数最少
- 等价于在有向图中搜索最短路径问题





最少匹配算法(Fewest Words Matching, FWM))

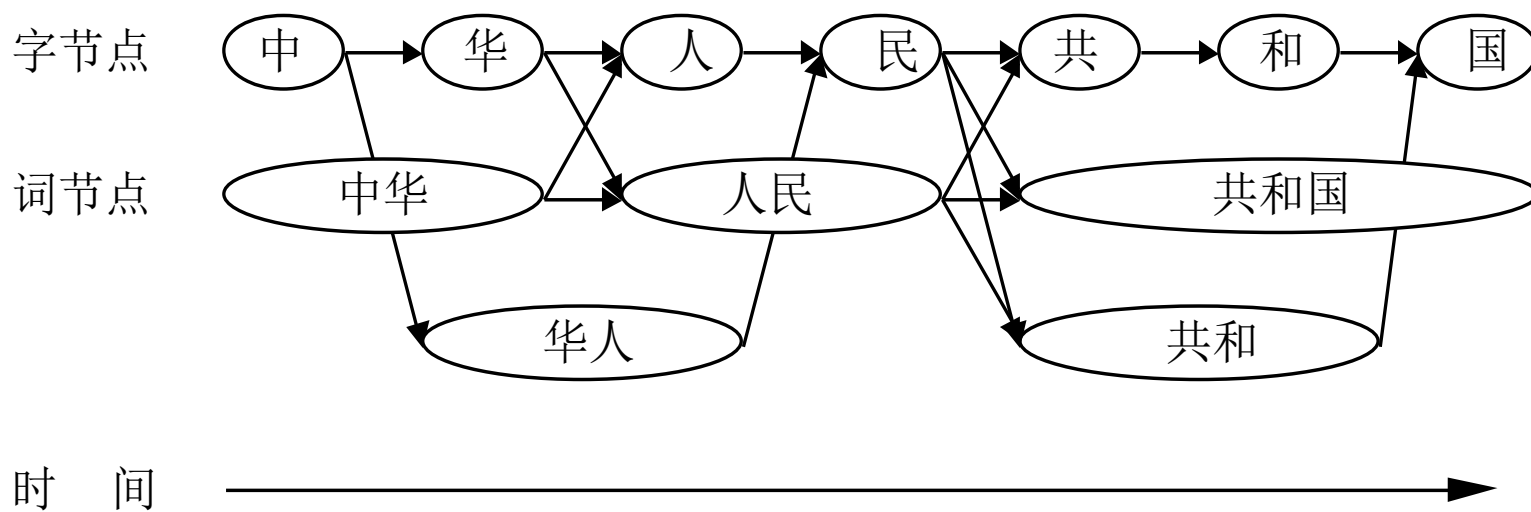
- 分段
- 逐段计算最短路径(Dijkstra算法)
- 得到若干分词结果
- 统计排歧
 - 发展\中\国家
 - 发展\中国\家
- 算法复杂性与FMM相当



基于统计的词网格分词

- 第一步是候选词网格构造：利用词典匹配，列举输入句子所有可能的切分词语，并以词网格形式保存
- 第二步计算词网格中的每一条路径的权值，权值通过计算图中每一个节点（每一个词）的一元统计概率和节点之间的二元统计概率的相关信息而得到
- 根据图搜索算法在图中找到一条权值最大的路径，作为最后的分词结果

字串“中华人民共和国”的切分词网格





分析

- 可利用不同的统计语言模型计算最优路径
- 具有比较高的分词正确率
- 算法时间、空间复杂性较高



生词识别

- 如何度量生词？
- 如何识别不同类别生词？
- 有通用的方法吧？



一种基于N-gram信息的生词获取

- 基本思想：N元对→词频过滤→互信息过滤→校正→生词获取

- 词频

- 互信息 (Mutual Information)

$$I(w_1; w_2) = \log \frac{p(w_1, w_2)}{p(w_1) \times p(w_2)}$$

- 词频与互信息的关系

- 候选生词的校正



一些抽取出的新词（三元组）

字数	抽取出的新词
3	阿拉伯（地名）、艾滋病、白求恩（人名）、独联体（组织名）、洞庭湖（地名）、工商局（机构名）、摄氏度（计量单位）、世乒赛（缩略名）、塔利班（组织名）
4	标本 兼 治（成语）、求 真 务实、萨 马兰 奇（人名）、神 州 大地、升旗 仪式、体制 转 轨、政企分开、通 货 膨胀（术语）、玩 忽 职守、新闻 媒 体、音 像 制品、优胜 劣 汰
5	奥地利 先 令（货币名）、波 黑 穆斯林（地名）、抽样 合格 率（术语）、电视 连续 剧
6	反 法西斯 战争、高 新技术 产业、工商 行政 管理、股份 有限 公司、国民 生产 总值（术语）
7	农村 剩余 劳动力、全国 人大 常委会（机构名）、香港 特别 行政区（地名）、常驻 联合国 代表



一些抽取出的新词（二元组）

字数	抽取出的新词
2	芭蕾、搬迁、北约（组织缩略名）、波黑（地名）、车臣（地名）、扶贫、乔石（人名）、印度（地名）、空调、欧盟（组织缩略名）、环保、媒体、拚搏、研讨
3	菜 篮子、反应 堆、党 组织、房 地产、副 主席（职位名）、国库 券、核 电站、价值 观、乒乓 球、食用 菌、实验 室、市 政府（机构名）、舒 马赫（人名）、消费 者、许可 证
4	百货 大楼、博士 学位、长篇 小说、犯罪 分子、改革 开放、高速 公路、国有 资产、绿色 食品、外汇 储备、知识 产权
5	供销 合作社（机构名）、天安门 广场（地名）、珠江 三角洲（地名）、最惠国 待遇、博士生 导师（职位名）、赤道 几内亚（地名）、钢筋 混凝土、三军 仪仗队、唯物 辩证法
6	辩证 唯物主义、工农业 总产值、国务院 副总理（职位名）、外交部 发言人、义勇军 进行曲、犹太人 定居点、计划经济 体制、联合国 安理会（机构名）、内蒙古 自治区（地名）
7	劳动人民 文化宫、塞尔维亚 共和国（地名）、无产阶级 革命家、中共中央 政治局（机构名）



生词的其他统计特征

- 统计构词能力
- 汉字构词模式
- 字对的亲合力



统计构词能力

$$WFP(c) = \frac{Count(\text{含}c\text{的多字词})}{Count(c)}$$

$$P_{WFP}(w) = \begin{cases} 1 - WFP(c), & |C|=1 (w \text{ 是单字词}) \\ \prod_{c_i \in w} WFP(c_i), & |C| > 1 (w \text{ 是多字词}) \end{cases}$$



汉字构词模式

$$P_r(pttn(c) | c) = \frac{Count(pttn(c))}{Count(c \text{ 位于多字词})}$$

$$P_{pttn}(w) = \prod_{i=1}^l P_r(pttn(c_i) | c_i)$$



字对的亲合力

$$P_r(t(c_i c_{i+1}) = t_B \mid c_i c_{i+1})$$

$$P_r(t(c_i c_{i+1}) = t_N \mid c_i c_{i+1})$$



人名识别

- 规则方法：利用语言规则来进行人名识别。优点：识别较准确；缺点：很难列举所有规则，规则之间往往会顾此失彼，产生冲突，系统庞大、复杂，耗费资源多但效率却不高
- 统计方法：一种是仅从字、词本身来考虑，通过计算字、词作人名用的概率来实现，另一种结合基于统计的汉语词语边界划分来实现。统计方法占用的资源少、速度快、效率高，但准确率较低。其合理性、科学性及其所用统计源的可靠性、代表性、合理性难以保证。搜集合理的有代表性的统计源的工作本身也较难。
- 混合方法：取长补短



一种基于统计和规则的人名识别方法

- 中文姓名用字特点（82年人口普查结果）
 - 729个姓氏用字
 - 姓氏分布很不均匀，但相对集中
 - 有些姓氏可用作单字词
 - 名字用字分布较姓氏要平缓、分散
 - 名字用字涉及范围广
 - 某些汉字既可用作姓氏，又可用作名字用字



人名识别系统资源

- 语料库：95、96两年的人民日报语料全集。共约**4000**万字。
- 人名库：包含共约**31000**多个人名。是95、96两年人民日报语料的所有人名的集合。
- 人名库和语料库的一致性对保证统计数据的准确性至关重要。



人名识别系统知识库

- 姓氏用字频率库和名字用字频率库：653个单姓氏，15个复姓，1894个名字用字

$$p(c\text{作为姓氏}) = \frac{c\text{用作姓氏的次数}}{c\text{的总出现次数}}$$

$$p(c\text{作为名字用字}) = \frac{c\text{用作名字用字的次数}}{c\text{的总出现次数}}$$



人名识别系统知识库

■ 名字常用词表

朝阳	劲松	爱国
建国	立新	黎明
宏伟	朝晖	向阳
海燕	爱民	凤山
雪松	新民	剑峰
建军	红旗	光明



人名识别系统知识库

■ 称谓库

■ 三种类型

- 只能用于姓名之前，如：战士、歌星、演员等；
- 只能用于姓名之后，如：阁下、之流等；
- 姓名前后皆可，如：先生、主席、市长等。

■ 称谓前缀表：“副”、“总”、“代”、“代理”、“助理”、“常务”、“名誉”、“荣誉”等



人名识别系统知识库

- 简单上下文

- 指界词表：约110个词

- 动词：说、是、指出、认为、表示、参加等；
 - 介词：在、之、的、被、以等；
 - 正在、今天、本人、先后等。

- 标点符号集

- 人名出现在句首或句尾（包括分句）的机会比较大，标点符号可用来帮助判断人名的边界。
 - 顿号一边是人名时，另一边的候选人名的可靠性高。



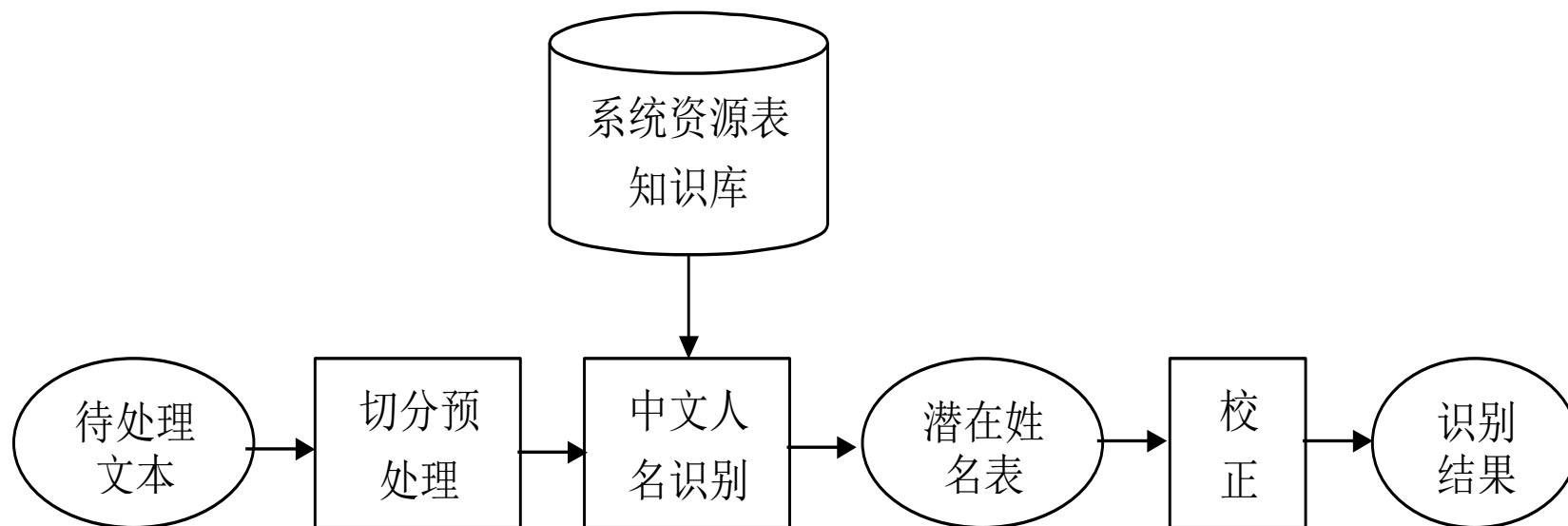
人名识别系统知识库

- 非名字用词表：有些双字词，如：时间、奖励、纬度等不作名字用词，但因为组成它们的单字可作为名字用字，如果跟在姓氏后面，往往会将其与可作姓氏的字一起误判为姓名。

例：

“做\这\件\事\花\了\我们\一\段\时间\
\”

中文人名识别过程



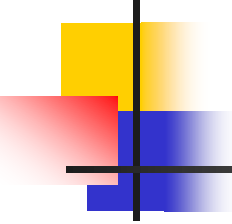


人名识别的具体实现

- → 姓氏判别
- → 名字识别
- → 概率判断

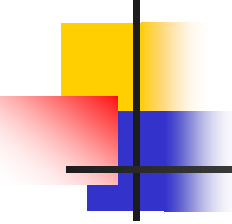
候选字符串为人名的概率为：

$$P = \text{姓氏部分为姓氏的概率} P_1 * \text{余下部分的汉字作名字用字的概率} P_2 * P_3 \text{ (单名时, 为} P_2 \text{)}$$



校正(对潜在人名的后处理)

- 当两个已辨识的人名相似时，需要检查是否要更正
 - C1C2C3与C1C2C4同时存在，C1C2正确；
 - C1C2C3与C1C2C4同时存在，C1C2C3正确；
 - C1C2C3与C1C2同时存在，C1C2正确；
 - C1C2C3与C1C2同时存在，C1C2C3正确



校正(对潜在人名的后处理)

- 自动校正:

- 如果两个潜在人名相似，考察它们的权值。
- 一高一低时，将低权值的潜在人名清除(李文常、李文)；
- 都为高权值时，两者都认为是人名(刘文军、刘文俊)；
- 都是低权值时，则各自通过第三个字作名字用字的概率大小来判断。概率够高，识别为人名。否则将第三个字去掉(李文常、李文及)。

- 人工校正



人名识别结果与分析

- 实验结果：8个测试样本，共22000多字，共有中文人名270个。系统共识别出中文人名330个，其中267个为真正人名。

召回率=文本中的中文人名辨识正确的比例
 $=267/270*100\% =98.89\%$

准确率=真正辨识正确的人名的比例
 $=267/330*100\% =80.91\%$

准确率和召回率是互相制约的，可通过概率阈值的调整来调节二者的关系。



人名识别结果与分析

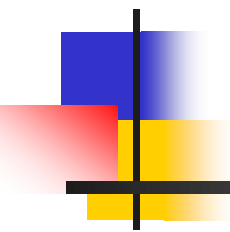
■ 产生错误的主要原因

- 被未识别的地名干扰。“湖北\英\山\县\詹\家\河\乡\陶\家\河\村\，\ ”
- 受非中式人名的干扰。“司\马\义\·\艾\买\提\ ”
- 分词结果不理想。“为\迎接\香港\回\归\送\贺\礼\”
- 规则不准确。“南\宋\大\诗人\杨\万\里\“\惊\如\汉\殿\三\千\女\，\ ”
- 其他。“全世界\每年\影片\产量\高\达\两\三\千\部\，\ ”



改进措施

- 采用更好的分词系统
- 构建更准确的姓名用字库、指界词库等
- 识别时结合一些语法、语义知识
- 采用更合理的大规模人名语料进行训练，使阈值确定得更合理
- 增加一些校正措施



谢谢！
