



OOV识别：未登录词

- ▶ 未登录词又称为生词，可以有两种解释：
 - ▶ 一是指已有的词表中没有收录的词。
 - ▶ 二是指已有的训练语料中未曾出现的词，此时未登录词又称为集外词。
- ▶ 通常情况下，将未登录词与OOV看作一回事。

OOV识别：未登录词分类

- 新出现的普通词汇，例如超女、博客、给力，以网络用语为主。
- 专有名词：包括人名、地名、组织机构名、时间、数字表达等。
- 专业名词和研究领域名称：特定领域的专业名词和新出现的研究领域名称也是造成生词的原因。
- 其他专有名词：新出现的产品名，电影书籍等文艺作品的名称。

OOV识别：一种基于N-gram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

OOV识别：一种基于N-gram的生词获取

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

OOV识别：一种基于N-gram的生词获取

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）

OOV识别：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、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

OOV识别：人名识别

- 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

OOV识别：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。

OOV识别：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

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

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

OOV识别：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

OOV识别：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

OOV识别：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。

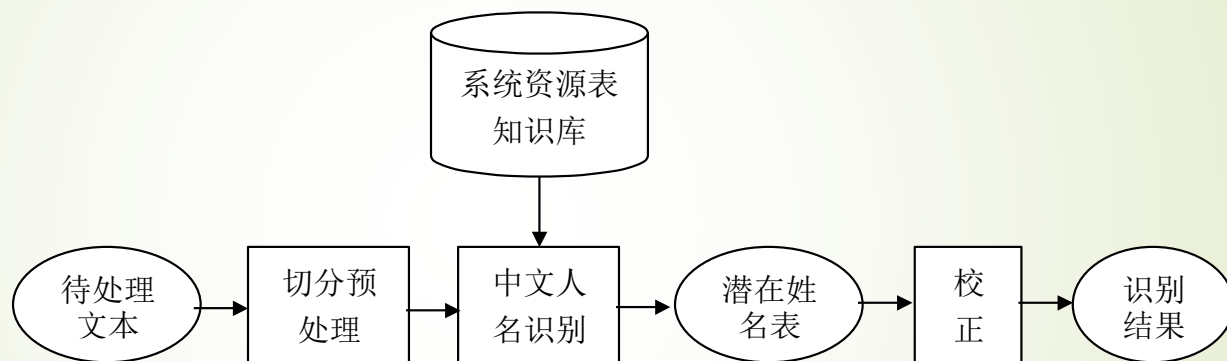
标点符号集

- 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。

OOV识别：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\。 \”

OOV识别：中文人名识别过程



OOV识别：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：

$P = \text{姓氏部分为姓氏的概率} P1 *$

$\text{余下部分的汉字作名字用字的概率} P2 *$

$P3 (\text{单名时, 为} P2)$

OOV识别：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

OOV识别：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

OOV识别：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。

OOV识别：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\, \ ”



OOV识别：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