第四章 未登录词获取

目录

未登陆词问题（1/4）

- 未登陆词主要包括两大类：
  - 新出现的词汇、短语或专业术语等，例如：博客、超女、恶搞、禽流感、裸退……
  - 人名、地名、组织机构名称等，例如：蔡国庆、张建国、新右卫门、山本五十六、约翰.斯特朗、詹姆斯.埃尔德、人民公园、中国科学院自动化研究所……

未登陆词问题（2/4）

- 以下是来自真实文本的例子：
  - 他还兼任何应钦在福州办的东路军军官学校的政治教官。
  - 林徽因此时已离开了那里。
  - 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。
  - 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。

未登陆词问题（3/4）

- 未登陆词的识别面临很多困难：
  - 由普通词汇构成，长度不定，也没有明显的边界标志词。
  - 专有名词的首词和尾词可能与上下文中的其他词汇存在交集型歧义切分。
  - 新出现的通用词汇和专业术语：面临词的界定问题。又回到了分词规范的问题上。

未登陆词问题（4/4）

- 在真实文本的切分中，未登陆词总数的大约九成是专有名词，其余的为通用新词或专业术语[黄昌宁等，2003]。
- 在自然语言处理研究中，人们通常将专有名词和数字、日期等词通称为命名实体（named entity）。命名实体识别NER是汉语自动分词研究中的关键问题之一。

02

基于统计的未登录词获取方法

基于统计的未登录词获取方法

- 基于频率的方法
- 基于均值和方差的方法
- 基于假设验证的方法
- 基于互信息的方法

基于频率的方法

- 简单计数

表 4.1 单纯依赖于频率的搭配发现。C(·)表示语料库中统计对象出现的频率

$C(W_1W_2)$	W_1	W_2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18566	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	YOAK
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

基于频率的方法

- 计数+词性过滤器（Justeson&Katz）

表 4.2 搭配过滤器的词性标记模式

标记模式	示例
A N	liner function
N N	regression coefficients
A A N	Gaussian random variable
A N N	cumulative distribution function
N A N	mean squared error
N N N	class probability function
N P N	degress of freedom

注：Justeson和Katz使用这些模式在频繁出现的词串中辨别可能的搭配

表4.3 搭配发现：Justeson和Katz的词性过滤器（P44）

表 4.3 搭配发现：Justeson 和 Katz 的词性过滤器

$C(W^1W^2)$	W^1	W^2	标记模式
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

基于均值和方差的方法

- Smadja提出的基于方差的词搭配方法
  计算语料中2个词之间的偏移量（有符号的距离）的均值和方差。寻找低偏差值的词对
  例：knocking on（或 at）the door
  a. she knocked on his door.
  b. they knocked at the door.
  c. 100 women knocked on Donaldson's door.
  d. a man knocked on the metal front door.

基于均值和方差的方法

上例中knocked和door之间的平均偏移量：

$$d = \frac{1}{4}(3+3+5+5)=4.0$$

利用公式 $s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$ 来估计方差。则此例偏差：

$$s = \sqrt{\frac{1}{3}((3-4.0)^2+(3-4.0)^2+(5-4.0)^2+(5-4.0)^2)} \approx 1.15$$

如果均值接近于1.0并且偏差较低，此类短语利用Justeson和Katz的基于频率的方法即可实现。如果均值远远大于1.0，那么一个低偏差预示了感兴趣的短语。

基于均值和方差的方法

表 4.5 基于均值和方差方法的搭配发现

s	d	计数	第一个单词	第二个单词
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organization
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

基于假设验证的方法

- t-test
- 卡方校验
- 似然比

基于假设验证的方法

- H_0:表明如果2个词不能形成一个搭配
  $P(w_1w_2)=P(w_1)P(w_2)$
- t-test

T-检验的基本原则是假定样本数据来自均值为 μ 的正态分布，然后通过对比样本均值和预期的均值 μ 之间的差异，判断样本是否来自于所假设的分布，从而推断出原假设是否成立。

基于假设验证的方法

- 根据T检验统计理论的原理，假设两词项x和y，在某语料库中共现概率为P(x,y)，各自单独出现的概率为P(x)和P(y)，那么所观测到的共现概率P(x, y)与随机共现的偶然概率P(x)P(y)之间的T值为:

$$T \equiv \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}} = \frac{p(x,y)-p(x)p(y)}{\sqrt{\frac{p(x,y)(1-p(x,y))}{N}}} \approx \frac{p(x,y)-p(x)p(y)}{\sqrt{\frac{p(x,y)}{N}}}$$

这里 $\bar{x}$ 为观察到的样本均值，s^2 为观察到的样本方差
N 为样本大小，μ 为假设分布的均值

- T-检验临界值(自由度N-1趋向于无穷大)：
  - T>1.96，假设成立的置信度< 0. 05
  - T>2.58，假设成立的置信度< 0.01
  - T>2.81，假设成立的置信度< 0. 005
  - T>3.29，假设成立的置信度< 0. 001

一般我们并不考虑置信度，仅仅实用T值来排序

stefanie

T检验（t-test）

- 在语料库中，new出现15828次，companies出现4675次，总共14307668个词。

零假设，new和companies独立出现

H_0: P(new companies)= P(new)P(companies)

$\approx 3.615\times10^{-7}$

$s^2=p(1-p)\approx p$ 这个分布

在语料库中，new和companies出现了8次，样本均值：

$$\bar{x}=\frac{8}{14307668}\approx 5.591\times10^{-7}$$

$$t=\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}}=\frac{5.591\times10^{-7}-3.6125\times10^{-7}}{\sqrt{\frac{5.591\times10^{-7}}{14307668}}}\approx 0.999932<1.96$$

表4.5 搭配发现： 把t检验应用到出现频率为20的10个二元组上（P48）

t	$C(w_1)$	$C(w_2)$	$C(w_1w_2)$	w_1	w_2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	Videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

卡方校验

χ^2 （chi-square）

利用联立表（ contingency table ）

	Wt+	Wt-
Ws+	31,950(a)	12,004(b)
Ws-	4,793(c)	848,330(d)

$$\chi^2=\frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

- 卡方检验临界值：
  自由度 =(行数 -1)*(列数 -1)=1
  - X^2>3. 84 ， 假设成立的置信度< 0. 05
  - X^2>6. 63 ， 假设成立的置信度< 0. 01
  - X^2>7. 88 ， 假设成立的置信度< 0. 005
- T- 检验假设概率分布为正态分布，这通常并不成立， x^2 检验没有这个要求。
- 一般我们仅使用卡方值来排序，并不使用置信度

似然比

似然比：假设检验的另一种方法，更适用于稀疏数据，它是一个简单数据，告诉人们一种假设的可能性比其他假设大多少。

将似然比检验应用到搭配发现的过程中，一般用下面两个可选的假设来解释W_1W_2 的出现频率：

假设 1： $P(w_2|w_1)=p=P(w_2|\neg w_1)$

假设 2： $P(w_2|w_1)=p_1\neq p_2=P(w_2|\neg w_1)$

假设1为独立性假设，假设2为非独立性假设。

使用最大似然估计的方法计算p,p_1和p_2，用c_1,c_2和c_{12}来表示在语料库w_1,w_2和w_{12}出现的次数，则：

$$p=\frac{c_2}{N}\quad p_1=\frac{c_{12}}{c_1}\quad p_2=\frac{c_2-c_{12}}{N-c_1}$$

似然比

$$b(k;n,x)=\binom{n}{k}x^k(1-x)^{(n-k)}.$$

表4.9 如何计算Dunning的似然比检验

	H_1	H_2
$P(w^2\|w^1)$	$p=\frac{c_2}{N}$	$p_1=\frac{c_{12}}{c_1}$
$P(w^2\|\neg w^1)$	$p=\frac{c_2}{N}$	$p_2=\frac{c_2-c_{12}}{N-c_1}$
条件c_1下c_{12}对应的二元组是 w^1w^2	$b(c_{12};c_1,p)$	$b(c_{12};c_1,p_1)$
条件$N-c_1$下c_2-c_{12}对应的二元组是 $\neg w^1w^2$	$b(c_2-c_{12};N-c_1,p)$	$b(c_2-c_{12};N-c_1,p_2)$

实际观测到的w_1，w_2和w_1w_2频率的似然值L(H1)(假设1的情况)和L(H2)(假设2的情况)是最后两行的乘积．即：

$L(H_1)=b(c_{12};c_1,p)b(c_2-c_{12};N-c_1,p)$

$L(H_2)=b(c_{12};c_1,p_1)b(c_2-c_{12};N-c_1,p_2)$

$$\log\lambda = \log\frac{L(H_1)}{L(H_2)}$$

并且他们分别对应于w_1w_2和$\neg(w_1w_2)$的出现频率。

总结

- 互信息:对于稀疏数据存在过高评价
- 卡方测试:数据稀疏情况下表现不好
- 对数似然比:可以识别稀疏二元搭配

stefanie

03

命名实体识别

命名实体识别

- NER又称作专名识别，是自然语言处理中的一项基础任务，应用范围非常广泛。命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。NER包含以下model：
  - 3 class model : Location, Person, Organization
  - 4 class model : Location, Person, Organization, Misc
  - 7 class model : Time, Location, Organization, Person, Money, Percent, Date

小明 在 中南大学 的 鸟巢 看了
PER ORG LOC
中国男篮 的一场比赛
ORG

命名实体识别

- **NER系统就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体，**比如产品名称、型号、价格等。因此实体这个概念可以很广，只要是业务需要的特殊文本片段都可以称为实体。命名实体识别技术是信息抽取、信息检索、知识图谱、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。

命名实体识别

- 在进行实体识别的过程中，有两个问题是十分关键的：
  - 实体边界的确认

是对一个句子中的实体词进行正确的划分。例如在句子“小明 在 北京大学 的 燕园 看了”中，一个好的识别算法必须将实体词进行正确的标记，而不是在其他的位置进行划分

  - 实体类别的判断

命名实体识别

- 序列标注

腰部 疼痛 加重 3 年

B-身体部位 B-医学发现 I-医学发现 B-时间词 I-时间词

命名实体识别

- 序列标注

$$f: X \rightarrow Y$$

Sequence Sequence

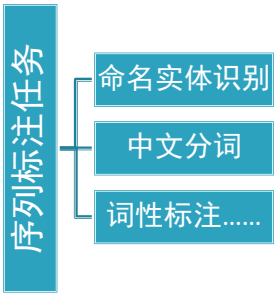
x: x_1 x_2 x_3 …… x_L (sequence)

y: y_1 y_2 y_3 …… y_L (sequence)

序列标注

- 序列标注任务的一般形式：

$X=\{x_1, x_2, \ldots, x_n\}$ ⇒ $Y=\{y_1, y_2, \ldots, y_n\}$

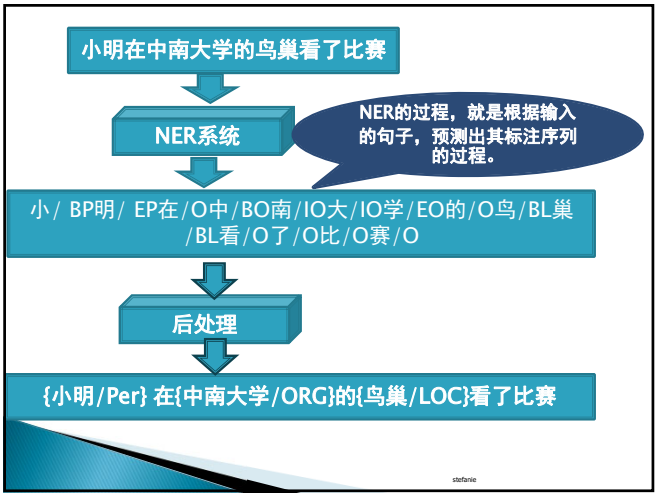


命名实体识别的数据标注方式

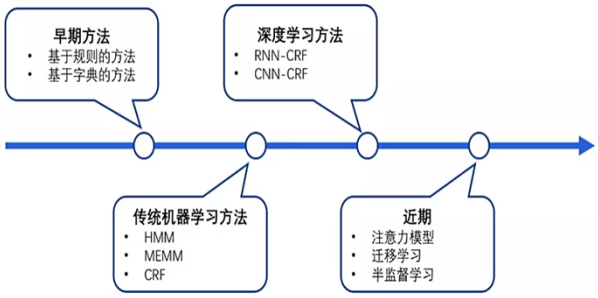
- NER是一种序列标注问题，因此他们的数据标注方式也遵照序列标注问题的方式，主要是BIO和BIOES两种。这里直接介绍BIOES，明白了BIOES，BIO也就掌握了。
  - B，即Begin，表示开始
  - I，即Intermediate，表示中间
  - E，即End，表示结尾
  - S，即Single，表示单个字符
  - O，即Other，表示其他，用于标记无关字符

命名实体识别的数据标注方式

- 将“小明在中南大学的鸟巢看了中国男篮的一场比赛”这句话，进行标注，结果就是：
- [B-PER，E-PER，O，B-ORG，I-ORG，I-ORG，E-ORG，O，B-LOC，E-LOC，O，O，B-ORG，I-ORG，I-ORG，E-ORG，O，O，O，O]



命名实体识别的方法



参考论文《A Survey on Deep Learning for Named Entity Recognition》

命名实体识别的方法

模型		优点	缺点
ME	最大熵	通用性好	训练效率低
MEMM	最大熵马尔可夫模型	充分利用特征	局部最优
HMM	隐马尔可夫模型	训练快	局部最优
SVM	支持向量机	理论完备	训练效率低
CRF	条件随机场	特征灵活、全局最优	训练特征模板

基于深度学习的方法

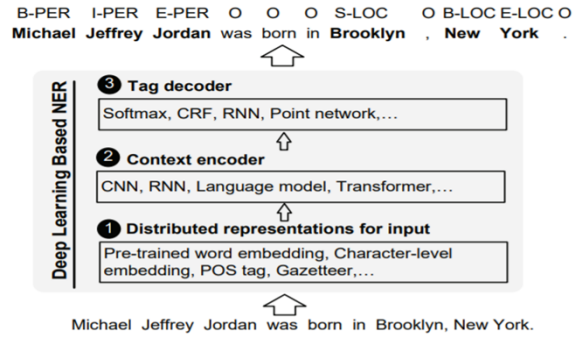
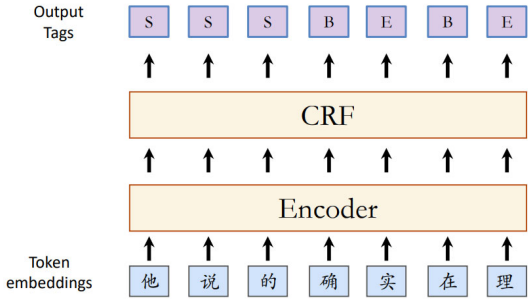


Fig. 2. The taxonomy of DL-based NER. From input sequence to predicted tags, a DL-based NER model consists of distributed representations for input, context encoder, and tag decoder.

序列标注



中文命名实体的工具

- **HanLP**
  https://github.com/hankcs/pyhanlp

1. HanLP

	算法	原理	特点
HanLP 命名实体识别	中国人名识别	HMM-Viterbi	速度
	地名识别		
	音译人名识别	层叠隐马	
	日本人名识别		
	机构名识别		
	感知机命名实体识别	感知机	精度
	CFR命名实体识别	CRF	

中文命名实体的工具

- **HanLP**
  https://github.com/hankcs/pyhanlp

1. HanLP

	算法	原理	特点
HanLP 命名实体识别	中国人名识别	HMM-Viterbi	速度
	地名识别		
	音译人名识别	层叠隐马	
	日本人名识别		
	机构名识别		
	感知机命名实体识别	感知机	精度
	CFR命名实体识别	CRF	

```
# HanLP1.7.7版本
from pyhanlp import *
from collections import Counter

with open("doc.txt","r",encoding="utf-8-sig") as f:
    txt = f.read()

# 中国人名识别
nlp=HanLP.newSegment().enableNameRecognize(True)

# 地名识别
# nlp = HanLP.newSegment().enablePlaceRecognize(True)

# 机构名识别
# nlp = HanLP.newSegment().enableOrganizationRecognize(True)

doc = nlp.seg(txt)
c = Counter()
for w in doc:
    if w.toString().find("nr")>=0:
        ww = w.toString()
        name = ww.split("/")[0]
        c[name] += 1
print(c.most_common(50))
```

➢ **HanLP实体标注：**

nr：人名

ns：地名

nt：机构名

➢ **执行结果：**

[('刘启林', 1)]

2. CRF++

开发：日本人Taku Kudo博士（C++）

实现算法：CRF模型

GitHub：https://github.com/taku910/crfpp

其它功能：Named Entity Recognition，Information Extraction，Text Chunking

命名实体的挑战

- 数量无穷
  - 不断发展,不断增加
- 构词灵活
  - 广州恒大淘宝俱乐部、广州恒大、恒大
- 类别模糊
  - 广州未赢够,广州下雪了

实体类别嵌套

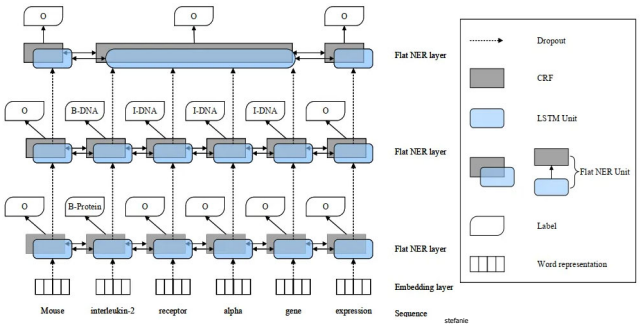
- 地名+组织=组织
  - (1)北京大学:【北京大学-组织】【北京-地名】
  - (2)上海市红十字会:【上海市红十字会-组织】【上海市-地名】【红十字会-组织】
- 人名+组织=组织
  - (1)王选计算机研究所:【王选计算机研究所-组织】【王选-人名】【计算机研究所-组织】
- 品牌+产品=产品
  - (1)兰蔻大粉水:【兰蔻大粉水-产品】【兰蔻-品牌】【大粉水-产品】
  - (2)华为mate38:【华为mate30-产品】【华为-品牌】【mate30-产品】

实体类别嵌套

- 4、地点+事件=事件
  - (1)上海婚博会:【上海婚博会-事件】
- 5、人名&作品
  - (1)老舍文集:【老舍文集-作品】【老舍-人名】
  - (2)电影刘三姐的剧情和内容非常的感人肺腑:结合上下文【刘三姐-作品】
- 6、事件&时间
  - (1)二战前夕:“二战”标“事件”,“二战前夕”标“时间”
- 7、地点&地点
  - (1)北京西站:“北京”和“北京西站”标“地点”

解决嵌套命名实体识别任务的经典模型

- 层叠式模型（Layered-based Model），该模型是通过堆叠多个Flat NER识别层的方式来识别Nested NER中的层次化结构



04

中文姓名的自动辨识

中文姓名的自动辨识

- 辨识姓名中的当用资源
- 姓名辨识的过程
- 姓名左右边界的确定
- 同源对表、互斥对表及其操作
- 屏蔽与恢复
- 同源对表、互斥对表的规则校正
- 概率再筛选
- 中文姓名辨识系统

辨识姓名中的当用资源

- 中文姓名的用字规律
- 中文姓名的使用频率
- 姓名的上下文

辨识姓名中的当用资源

- **中文姓名的用字规律**

通过对174　900个中文姓名进行抽样综合统计的结果，建立姓氏频率表XFL与名字用字频率表MCFL。发现以下规律：

1. 统计的到姓氏并不多，仅为729个，分布很不均匀，但相对集中。前5大姓“王、陈、李、张、刘”占了姓名样本库的32%，前365个姓占99. 0%，其余364个姓氏仅占不到1%。
2. 某些姓氏可用做单字词，其中不乏高频单字词，如常见姓氏“王、黄、马、高、于”和不常见的姓氏“是、过、来、从、那”。
3. 某些汉字既可做姓氏，又可用作名字用字。如“林、方、金、江、柳”。

辨识姓名中的当用资源

4.统计得到的3345个用字。名字用字的分布较姓氏要平缓分散涉及的范围很广，从所属的词类上看，不仅有实词，还有各类虚词

5. 根据构词能力，名字用字可以划分成三类：即可用做单字词的开放式名字用字：虽不可用单字词，但可构词的相对封闭式名字用字：以及既不可用作单字词，有不可构词的绝对封闭式名字用字。例如“爱” 、“睿” 、“逵”就分别属于这三类字。

辨识姓名中的当用资源

- **中文姓名的概率分布**

用f(x)（x∈姓氏）表示姓氏x的使用频率；

用f(m)（m_i∈名字用字，i=1,2）表示名字用字m_i的使用频率。

根据XFL及MCFL，可给出姓名的概率估值：

$P(sn)=f(x)*f(m_1)$或者$P(sn)=f(x)*f(m_1)*f(m_2)$

计算姓名样本库中所有姓名的概率估值后发现，对数概率估值曲线呈陡峭的单峰分布，极高或极低的概率估值均不多，因此，可以设定概率估值阈值，舍弃那些概率估值小于阈值的候选名字。

辨识姓名中的当用资源

- **具有指示意义的上下文**

一些上下文信息有助于姓名的辨识，主要有称谓、指界动词 、匹配模式这些上下文信息和姓氏频率表XFL和名字用字频率表MCFL一起构成中文姓名辨识的知识源。

称谓常与名字同时出现，对姓名辨识有指示作用。例如：
“省长**赶到了救灾现场”称谓“省长”提示了姓名的左边界。
按照和名字的前后顺序，可以吧称谓分成三类：

1．只能用于姓名之后，如“之流”，“阁下”等。
2．只能用于姓名之前，如“青年”，“战士”等。
3．用于姓名前后均可，如“先生”，“市长”等。

一些动词，如“说、是、指出、认为”等，常常接在姓名的后面，可以用来帮助确定姓名的右边界。如“胡锦涛指出……”

姓名辨识的过程

输入文本分割成句子，并用最大匹配法分词之后，对句中的每个字加上标志。加标志主要根据当前字是否为孤立字，是否可做单字词，是否属于某个指界动词，是否属于某个称谓。

接下来寻找句中所有可能的潜在姓名cn，并添加到潜在姓名表CNL中。要求cn的姓氏用字在姓氏频率表XFL中，而名字用字在名字用字频率表MCFL中。在计算cn的概率估值，若cn的概率估值小于阈值，则舍弃之。但是如果cn的每个字都是孤立字，就要放宽阈值要求。

预切分、标志数组及概率初筛选算法设计如下：

姓名辨识的过程

- 输入文本分割成句子后送入数组SENT
- 在常用词表的支持下，利用最大匹配法对SENT进行分词，分词结果仍存回SENT。但保持SENT的长度不变。
- 根据上节所述资源，建立一个与经分词处理后的数组SENT具有一一对应关系的标志数组FLAG。

姓名辨识的过程

赋加标志遵循的规则是：

对句内任一位置pt（0<pt<length(SENT)-1）
若SENT[pt]为孤立字且该字不可作单字词
FLAG[pt] ←’0’
若SENT[pt]为孤立字且该字可作单字词
FLAG[pt] ←’1’
若SENT[pt]属于某个多字词(含双字词)
FLAG[pt] ←’2’
若SENT[pt]属于某个称谓
FLAG[pt] ←’3’
若SENT[pt]属于某个指界动词(含单字词)
FLAG[pt] ←’4’
若SENT[pt]属于某个指界动词(含双字词)
FLAG[pt] ←’5’
若SENT[pt]是分割符(数字、字母、标点等非汉字)
FLAG[pt] ←’6’
赋加标志的优先权：’4’>’1’;’3’>’5’>’2’

姓名辨识的过程

接着，算法将寻找SENT中所有可能的潜在姓名。设SENT中任一连续汉字串CSTER
C1C2：若有C1(单姓)∈XFL，C2∈MCFL；或 //单姓单名
C1C2C3：若有C1(单姓)∈XFL，C2，C3∈MCFL；或 //单姓双名
C1C2C3：若有C1C2(单姓)∈XFL，C3∈MCFL；或 //复姓单名
C1C2C3C4：若有C1C2(单姓)∈XFL，C3，C4∈MCFL //复姓双名
则CSTR即被视作一潜在姓名cn，并将之添加到潜在姓名表CNL中。

1、若cn所属各字对应的FLAG值均为‘0’ 、 ‘1’或者‘4’
则state(cn)=1；否则state(cn)=0，
2、下一步是概率初筛选：
对任一cn ∈CNL，state(cn)=1：
若 (cn为单姓单名且lg(p(cn))< ξ2)或若 (cn为单姓双名且lg(p(cn))< ξ3)
则从CNL中删除cn state(cn)=0 /*情形1*/

姓名辨识的过程

若 (cn为复姓单名且lg(p(cn))< ξ4)或若 (cn为复姓双名且lg(p(cn))< ξ5)
则从CNL中删除cn /*情形2*/
取： ξ2=-8.538 992； ξ3=-11.903 766
ξ4=-6.239 499； ξ5=-10.068 238

姓名左右边界的确定

设一个潜在姓名cn的语境为：

context(cn)=zl cn zr(zl,zr为汉字)

则有边界确定规则：

1.若(FLAG(zl)=‘3’)或(FLAG(zl)=‘6’)或(cn在句首)，则cn左边界确定，记作＃cn。

2.若(FLAG(zr)=‘3’)或(FLAG(zr)=‘5’)或(FLAG(zr)=‘6’)或(cn在句尾)或(cn为双名且(FLAG(zr)=‘4’)),则：cn的右边界确定，记作cn＃。

3.若(＃cn)且(cn＃)，则cn被确认，并从潜在姓名表CNL中删除，送入确认姓名表OKL中。

(注意：一个姓名被“肯定”，仅表示着在同其他姓名的竞争中“生存”下来，取得了继续保留在潜在姓名列表中的资格。“确认”则是百分之百的肯定。)

同源对表、互斥对表及其操作

同源对为以句内同一位置为姓氏起点的单名与双名。互斥对为以句内不同位置为姓氏起点，同时相互间又有交叉的两个姓名。同源对和互斥对体现了潜在姓名之间的相互制约关系。一个潜在姓名被肯定，则所有和它同源或互斥的潜在姓名都将从CNL中删除。

屏蔽与恢复

屏蔽与恢复均针对同源对。

- **屏蔽**

  屏蔽发生在同源对中双名的名字末位置之FLAG值为‘0’时，如：

  同源对[刘仲，刘仲藜]

  其中‘藜’的FLAG值为‘0’，屏蔽有效。“刘仲”被否定。

- **恢复**

  对某些双名，如果仅根据概率值，在初筛选阶段会被滤掉。但若满足条件：名字末字位置的FLAG值为‘0’且同源对中名单的概率值＞ξ2，则须予以保留(即恢复)，如：

  lg(p(邵逸夫))＜ξ3

“邵逸夫”应该被滤掉。然而‘夫’的FLAG值为‘0’且lg(p(邵逸夫))＞ξ2，

则应恢复“邵逸夫”，同时屏蔽“邵逸”。

同源对表、互斥对表的规则校正

[校正规则1]同源对右界否定规则

若同源对形如[z1z2z3#,z1z2]则否定z1z2;

[校正规则2]互斥对左界否定规则

若同源对形如<#z1z2z3,z2z3＞则否定z2z3;

[校正规则3]互斥对落单否定规则

设互斥对对形如<(cn1:PA+INTERSECT),(cn2:INTERSECT+PB) ＞

/*INTERSECT为cn1,cn2交叉部分，且不为空*/

若cn1之PA部分至少含一FLAG值为‘0’的字，则否定cn2，

若cn2之PB部分至少含一FLAG值为‘0’的字，则否定cn1；

[校正规则4]互斥对等长概率否定规则

设互斥对<cn1,cn2＞中length(cn1)=length(cn2);

若p(cn1)＞p(cn2)则否定cn2；

若p(cn1)＜p(cn2)则否定cn1；

同源对表、互斥对表的规则校正

[校正规则5]互斥不对等长概率否定规则

设互斥对<cn1,cn2＞中length(cn1) ≠length(cn2);

若lg(p(cn1))/length(cn1)＞lg(p(cn2))/length(cn2)则否定cn2；

若lg(p(cn1))/length(cn1)＜lg(p(cn2))/length(cn2)则否定cn1；

/*关于lg(概率值几何平均)之比较*/

校正规则依所列顺序依次调用。校正规则5导致互斥表CTL终必变化为空。

概率再筛选

对经以上各步余下的潜在姓名表CNL，再进行一轮筛选。新的阈值为：

ξ2=-7.904 649; ξ3=-11.072 278

ξ2=-5.129 399; ξ3=- 8.881 645

仅视乎此阈值，召回率也将不低于98.69%。但由于透过多层次处理后，OKL表中已有相当积累，故辨识系统的实际召回率也比这个指标高。

中文姓名辨识系统

以清华大学孙松茂、黄昌宁等开发的中文姓名辨识系统为例，可以看到典型
的r

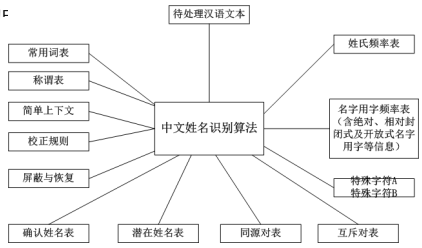


图4.1：中文姓名识别系统结构

中文姓名辨识系统

给出一个分析示例：
设输入句子为： 湖北柳大华和广东吕钦
则算法过程如下：
1. 经最大匹配切分后，得标志数组：
SENT :湖北柳大华和广东吕钦，
FLAG :2 2 1 1 0 1 2 2 0 0 6
给出潜在的姓名表CNL：
（柳大华-7.523170）（大华和-9.348916）（大华-6.632542）
（华和广-8.837197）（华和-6.101831）（和广东-8.609877）
（和广-6.200005）（广东吕-11.244786）（广东-6.807546）
（东吕钦-10.975300）（东吕-7.771527）（吕钦-5.019326）
这里，潜在姓名“柳大”已被“柳大华”所屏蔽。

中文姓名辨识系统

2. 概率初筛选后，CNL变成：
（柳大华-7.523170）（大华和-9.348916）（大华-6.632542）
（华和广-8.837197）（华和-6.101831）（和广东-8.609877）
（和广-6.200005）（吕钦-5.019326）
互斥对表CTL为：
<柳大华，大华和><柳大华，大华><柳大华， 华和广>
<大华和，华和广><大华，华和广><柳大华，华和>
<大华和，华和><大华，华和><大华和，和广东><华和广，和广东>
<华和，和广东><大华和，和广><华和广，和广><华和，和广>
同源表对SSL为：
[大华，大华和][华和，华和广][和广，和广东]

中文姓名辨识系统

3.运用同源表对、互斥表对的校正规则：
由[校正规则4]互斥对<柳大华，大华和>否定“大华和”；
由[校正规则4]互斥对<柳大华， 华和广>否定“华和广”；
由[校正规则4]互斥对<大华，华和> 否定“华和”；
由[校正规则3]互斥对<华和广，和广东>否定“和广东”；
由[校正规则3]互斥对<大华和，和广>否定“和广”；
注意，每一次运用规则后，CNL，CTL和SSL都要作相应调整。
此时有：
CNL（柳大华-7.523170）（华和-6.101831）（吕钦-5.019326）
CTL <柳大华，华和>
SSL 空表

中文姓名辨识系统

4.进一步，由[校正规则5]对互斥对<柳大华，华和>否定“华和”
5.最总结果是：
CNL（柳大华-7.523170）（吕钦-5.019326）
CTL空表
SSL 空表