基于混合策略的汉语未登录词整体识别*

于传武 李生 陈鄞 赵铁军

哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001

E-mail: ycw, lisheng, chenyin, tjzhao@mtlab.hit.edu.cn

摘要： 未登录词的识别一直是汉语分词研究的焦点和难点，本文通过对各类未登录词的用字频率及上下文进行了详细地分析，提出一种基于混合策略的未登录词识别方法。实验表明，该方法对于多种未登录词的整体识别取得了较好的结果。

关键词: 未登录词；汉语分词；混合策略；全局统计

引言

未登录词识别是影响汉语分词精度的一个主要因素。所谓未登录词是指分词系统词典中没有收录的词。汉语词汇是一个开放集合，无论建立多么庞大的词典，都不可能穷举所有的词。而且，随着时间的推移，还会源源不断地出现大量的新词。而且无限度往字典里加入新词，一方面会使系统的资源过于庞大，增加系统负担，降低运行效率；另一方面，收录的未登录词会和上下文产生冲突。比如句子“李鹏强考上了大学”，如果把“李鹏”作为人名收入字典，李鹏强就不能被识别出来，因此研究未登录识的自动识别就成为一个急需解决的问题。

目前，对于各类未登录词，如人名、地名、译名的识别，许多学者进行了大量的研究[1][3][4][5]，并且取得了不错的效果。从他们识别的方法看，大部分都是事先针对各类未登录词建立语料库，统计其用字频率，然后结合上下文，设立评价函数，并设定阀值，将高于阀值的作为未登录词识别出来。但这些方面也存在着一些不足，比如这些方法在处理单一未登录方面可能精确率比较高，但是很容易把某一类型的未登录词识别为另一类。本文所采用的识别方法是：首先要从识别文本中筛选出所有可能的候选词，设立统一的评价函数，对每个候选词用评价函数进行打分，然后针对各种可能用动态规则选出一种最好的结果。本文识别的未登录词主要包括中国人名、地名、外国译名。

1 未登录词的识别

1．各类未登录的统计信息

本文识别所用的统计资源包括中国人名库，含人名约五百万条；地名库，含地名库约五百万条；译名库，含译名库约五万条。通过对这些资源中用字进行统计，得出的统计信息如图 1 所示：

*本课题得到国家“863”项目基金(项目编号 2001AA114101)的资助

表 1 汉字用字统计信息表

	人名	地名	译名
首字（个）	1433	1907	297
中间字（个）	2626	1492	320
尾字（个）	3048	1681	259

由于人名和地名库规模比较大，译名库的规模比较小，在处理过程中把人名和地名中出现次数低于 10 次的自动剔除，把译名中出现次数低于 3 次的自动剔除。

从上表可能看出，人名、地名、译名的用字还是比较有规律的，特别是译名，只覆盖了所有汉字的 10%左右，这也说明用统计字频的方法处理汉语未登录词是可行的。图 1、图 2、图 3 分别给出了从语料库中统计的人名、地名、译名用字分布信息。图中横坐标表示汉字的个数，纵坐标表示对应汉字的分布。由图可以看出，人名、地名、译名的用字基本上都集中在一定范围内，这就给进行识别提供了一种可以遵循的方法。

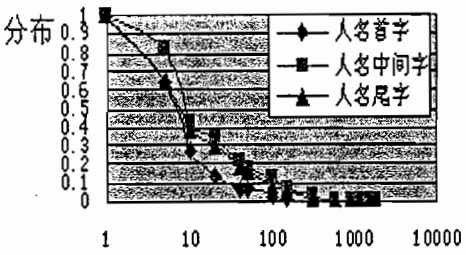

图 1 人名用字统计信息

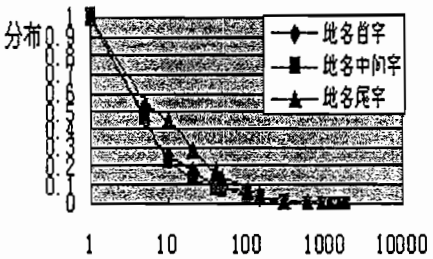

图 2 地名用字统计信息

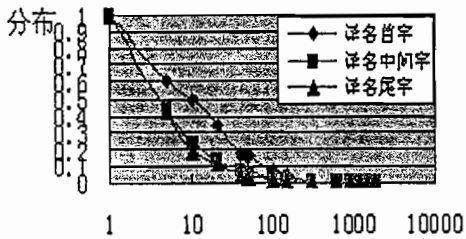

图 3 译名用字统计信息

为了便于各类未登录词用字信息的比较，我们统计了 6763 个常用汉字在不同类型用法中的可信度。可信度的计算公式如下：

$P_{nf}(c) = -\ln(N_{nf}(c)/N_{nfmax})$　$P_{nm}(C) = -\ln(N_{nm}(c)/N_{nmmax})$　$P_{nl}(C) = -\ln(N_{nl}(c)/N_{nlmax})$

$P_{lf}(c) = -\ln(N_{lf}(c)/N_{lfmax})$　$P_{lm}(C) = -\ln(N_{lm}(c)/N_{lmmax})$　$P_{ll}(C) = -\ln(N_{ll}(c)/N_{llmax})$

$P_{tf}(c) = -\ln(N_{tf}(c)/N_{tfmax})$　$P_{tm}(C) = -\ln(N_{tm}(c)/N_{tmmax})$　$P_{tl}(C) = -\ln(N_{tl}(c)/N_{tlmax})$

$P_g(c) = -\ln(N_g(c)/N_{gmax})$

其中，$P_{nf}(c)$，$P_{nm}(C)$，$P_{nl}(C)$分别表示汉字 c 做人名首字，中间字，尾字的可信度。$N_{nf}(c)$，$N_{nm}(c)$，$N_{nl}(c)$分别表示汉字 c 在人名语料库中做首字，中间字，尾字的频率。N_{nfmax}，N_{nmmax}，N_{nlmax}分别表示做人名首字，中间字，尾字最高频率汉字的频率。

公式中的第二行，第三行分别表示地名和译名的各种值，具体表示同人名表示相同。

$P_g(c)$表示汉字 C 做一般字的可信度，$N_g(c)$表示汉字 c 在统计一般词语料库中出现的频率，N_{gmax}表示最高频率汉字的频率。

为了使各类未登录词之间能相互比较，我们把算出来的可信度按照它们各自的分布分为六个等级，等级最低的表示它不能做该种类型的未登录词，等级最高的表示做该种类型未登录词的概率最大。

同时在人名、地名、译名中可能存在着大量的字典中的词，为了便于识别，还对人名、地名、译名语料库中常出现的词也分别在字典中做了标记，并区分该词在某种类型的未登录中是做首词、中间词还是尾词。

2. 寻找候选的未登录词

整个未登录的识别都是在分词过程中进行的，识别过程中，先对整句话进行全切分，把切分的结果用一个有向图表示出来，然后在这个有向图里寻找各种类型的候选未登录词，设该候选未登录词对应的字序列为 $f_i f_{i+1} \cdots f_j (j>i)$，则候选未登录必须满足以下条件：

（1） 如果该词是一个候选人名，则满足：$P_{nf}(word(f_i))>0$，$P_{nm}(word(f_m))>0 (i<m<j)$，$P_{nl}(word(f_j))>0$。

（2） 如果该词是一个候选地名，则满足：$P_{lf}(word(f_i))>0$，$P_{lm}(word(f_m))>0 (i<m<j)$，$P_{ll}(word(f_j))>0$

（3） 如果该词是一个候选译名，则满足：$P_{tf}(word(f_i))>0$，$P_{tm}(word(f_m))>0 (i<m<j)$，$P_{tl}(word(f_j))>0$

（4） 如果 $f_i f_{i+1} \cdots f_j$ 中有词，则该词必须是字典里已经收录能做该类型未登录词的词，并且满足做首词、中间词或者尾词的条件。

（5） 为了避免发生歧义，对于下列这种情况不在这个阶段进行识别：如果 $i>0$ 而且 f_i 和 f_{i-1} 组成词，或者 f_j 和 f_{i+1} 组成词。比如“张玉普通过了考试”，在这个阶段，只可能把张玉做为人名识别出来，而张玉普不进行识别，因为“普”和后面的“通过”有歧义出现，对这种情况，将放在分词后的后处理中进行识别。

对于地名和译名，由于它们长度的不确定性，这就需要对它们的边界进行识别。具体的策略是，首先，从已标注的语料库里找出地名和译名之前和之后常用词或者词性，比如外国人名之后会经常出现如：“说，指出，参加”等词，前面经常出现如“市长，首相，要求”等词，还有就是，人名经常会和一些称呼词连在一起使用，地名后面会和一些表示行政单位的词连在一起使用，译名中可能会含有“·”等等，对这些常用的和各类未登录词能搭配的词和词性进行收录，在识别时，如果候选的地名或译名符合上下文边界条件，就把它们作为候选词，否则进行以下处理：

设侯序词的字序列仍为 $f_i f_{i+1} \cdots f_j (j>i)$，如果 f_i 做未登录词首字的可信度小于做一般词的可信度，而且该字做未登词首字的可信度比较低，就继续向右扫描，直到找到一个位置 m，使 m 做未登录首字的可信度大于做一般词的可信度。然后按照同样的方法从右往左扫描，找到一个位置 n， 使 n 做未登录尾字的可信度大于做一般词的可信度。如果 $n>m$ 就把该词作为一个候选未登录词。

3. 规则的组织

通过词频得到的未登录词，只是考虑的词频，但在有些情况下，通过词频得到的候选未登录词在特定的上下文是不能作为未登录词，这就需要用规则的形式对其进行剔除，或者减少它作为未登录词的概

率。最后的评价函数实际上包括两部分：词频的可信度+上下文可信度。

上下文的可信度，包括奖励和惩罚两种可信度。奖励主要是从已进行分词和词性标注的语料中抽出能和各类未登录词进行搭配的常用词或词性，并给予相应的奖励。惩罚主要是指所识别的候选词在特定上下文中不可能作为未登录词，或者说作为未登录词的概率很低，就给与相应的惩罚。

下面列举了一些规则库中的常用规则：

人名奖励规则： 1:Cate=nc -> 4

1:W=本人-> 2

-1:W=与 ->2

人名惩罚规则： (2)0:(F)Cate=d&&(L)Cate=vz->-3

(0)-1:W=、+1:W=、->2

(3)0:(M)Cate=m&&(L)Cate=q->-8

地名奖励规则： +1:W=当局->2

+1:Cate=nq->4

地名惩罚规则： (2)0:(F)Cate=a&&(L)Cate=ng&&(L)W=寨->-20

(0)-1:Cate=m+0:(F)Cate=q->-4

上述奖励规则里冒号前面的0表示当前节点，1表示当前节点的下一个节点，-1表示上一个节点，Cate表示词性，W表示字。

对于惩罚规则如“ (2)0:(F)Cate=d&&(L)Cate=vz->-3”，最前头括号里的2表示当前节点由2个字组成，如果括号里为0表示是任意长度。(F)表示当前节点的首字，(L)表示当前节点的尾字，(M)表示当前节点的中间字。冒号前面的数字，及Cate，W的意思表示同奖励规则。该规则表示如果当前节点的长度是2，而且首字的词性是d，尾字的词性是vz，就把它的可信度减3。

通过这些规则也能把一些人名中的兼类情况进行处理，比如“张”字，在汉语里即可以做姓氏，又可以做量词，在做量词的时候他前面一般跟数词或者代词，通过规则(0)-1:Cate=r+0:(F)W=张->-2，(0)-1:Cate=m+0:(F)W=张->-2 这两条规则就能有效地降低“张”字在这种环境下做姓氏的机率。

4. 全局统计

在一篇文章中，有些未登录会反复的出现，在新闻领域中，这种现象尤为突出，这就给识别未登录词提供了一些重要的信息，因此，在识别的过程中加入全局统计就显得特别重要，所谓全局统计，就是任给一个词，就能从全文中自动搜出该词在整个文章中出现的次数，位置等等。对于全局统计来说，最重要的一个特点是要快，查找迅速，我们不能只是用简单的串匹配进行查找，因为这样效率特别低，会对整个系统的运行速度产生重大影响。

本文在处理的过程中采用的策略用邻接表的形式来存储该文章中某字在文章中的具体位置，首先根据汉字建立哈希表，每行存储的是该字在文章中所处的位置。在查找时只需根据所在查找词的首字用哈希函数进行定位，就能快速查找到该词在原文中出现的次数。

5. 动态规则

有了全局统计，对未登录的评价函数就改为 E(w)=词频的可信度+上下文可信度+全局统计的可信度。例如，给定句子“黔南镇党委书记聂瑞刚便带领工作人员到罗切斯特的家里参观”中的各种未登录词的评价函数值为：

表 2 各类未登录的评价值

类型	候选词及评价值
人名	E(南镇)=4　E(聂瑞)=13　E(聂瑞刚)=19　E(刚便)=-13 E(罗切斯)=6　E(斯特)=-4
地名	E(黔南)=11
译名	E(罗切斯特)=20
一般字	E(黔)=2　E(南)=4　E(镇)=3　E(聂)=1　E(瑞)=3　E(刚)=3　E(便)=4 E(罗)=3　E(切)=4　E(斯)=4　E(特)=4

在动态规划的过程中，其实规划的只是含有未登录的节点，对其余的节点，只是累加其做为一般词的可信度，通过动态规则，就能先把句子里的未登录词先确定出来，最终结果为：

/黔南/nd 镇党委书记/聂/nx /瑞刚/nm 便带领工作人员到/罗切斯特/ny 的家里参观。

6. 后处理

在分词过程中，一个难点就是未登录词和词典中其它的词出现歧义，比如“阿里斯特工业非常发达”。在这句话里，正确的切分形式应该是：阿里斯特/工业/非常/发达/。但由于“阿里斯特”和后面的“工业”产生了歧义，这就给处理增加了难度，如果在分词过程中直接按照未登录词的筛选方法进行处理，就可能对系统的歧义切分精度产生影响。对这种问题的研究我们目前还处于摸索阶段，主要是放在分词后处理中进行。这就要求系统具有良好的歧义处理能力，如果系统能把“特工业”按照“特/工业”的形式，系统在分词后面就会对每一个有歧义的未登录的后一个字进行处理，如果可能作未登录词末字，就把它跟前一个节点合并起来。

“阿里斯特工业非常发达”在分词阶段划分的形式为：阿里斯/特/工业/非常/发达/，经过后处理就会把“特”加入到前面的“阿里斯”节点中去，最终的切分形式为：阿里斯特/工业/非常/发达/。

2 实验结果

对于从 1998 年的人民日报语料中随机抽出一千句进行开放测试，所测得的数据及计算后的精确率和召回率如表 3 所示：

表 3 单句的测试结果

	总正确数	识别总数	正确识别	精确率	召回率

人名	1601	1550	1403	90.05%	87.6%
地名	1877	1824	1602	87.8%	85.3%
译名	384	343	323	94.5%	84.1%

表 4 篇章级的测试结果

	总正确数	识别数(单)	正确识别（单）	识别数（篇）	正确识别（篇）
人名	52	49	44	52	48
地名	58	53	48	57	50
译名	78	73	70	75	73

以上测试结果主要是针对单句进行测试的，我们又从“新浪”网站随机抽取几篇含人名地名比较多文章进行篇章级的测试，结果如表 4 所示。从测试的数据可以看出，篇章级的识别确实比单句识别的效果要好。例如，如果文章中反复出现“李铁强”，最后肯定会把“李铁强”识别出来，而不会出现只识别出“李铁”这种情况。但有时会增加识别的个数，主要原因是一些候选但最终不是未登录词的在句中反复出现，比如在新浪上测试的一篇体育文章中“赛前”出现了 8 次，结果系统最后就把他作为人名识别出来了，这些将是今后研究中需要注意的地方。

3 结论

在中文信息处理中，汉语的未登录词识别一直是一项比较基础，但又非常重要的研究工作，它识别的好坏对于汉语的自动分词，自动文摘，信息抽取，及至整个机器翻译都将产生深远的影响，本文尝试着用基于混合策略的方法来进行汉语未登录词的整体识别，实验结果是令人满意的，但是仍有很大的上升空间，同时在实现的过程中也发现在分词阶段可供利用的信息太少，如“张刚和马路对面的小孩有什么关系？”，在这个句子中，如果仅利用字和词性的信息，很难判断“和”字是否是人名的一部分，如果能利用一些浅层的句法分析，分析一下“和”后面的中心语，就可能知道“和”字具体的词性。另一方面，有些未登录词常常是字典里词，这就给识别带来了很大的困难。这些将是今后研究中重点加以解决的问题。

参考文献:

[1] 沈达阳，基于统计和规则的汉语真实文本自动分词和词性标注系统的研究与实现。清华大学硕士学位论文。1996
[2] 沈达阳 孙茂松 黄昌宁 中国地名的自动辨识 中文信息学报，1995 年第 2 期
[3] 郑家恒 刘开瑛 自动分词系统中姓氏我名处理策略探讨 山西大学计算机科学系
[4] 季恒 罗振声 基于反比例模型和规则的中文姓名自动辨识系统 自然语言理解与机器翻译 2001
[5] 吕雅娟 赵铁军等基于分解与动态规划策略的汉语未登录词识别 中文信息学报 2001

作者简介： 于传武 男，硕士研究生，主要研究领域为自然语言处理、机器翻译。 李生 男，博士生导师 主要研究方向是机器翻译。陈鄞 女 硕士研究生 主要研究方向是自然语言处理。赵铁军 男 博士生导师 主要研究方向为自然语言处理、机器翻译。

A Hybrid Method for Identification of Various Types of Unknown Words

YU CHUANWU LI SHENG CHEN YIN ZHAO TIEJUN

School of computer science and technology, Harbin Institute of Technology Harbin 150001

E-mail: ycw, lisheng, chenyin, tjzhao @mtlab.hit.edu.cn

Abstract: Unknown word identification has always been a key and open problem for Chinese segmentation. On basis of a detailed analysis of the character frequency and the context for unknown words, a hybrid method is proposed. The experiment shows satisfactory result for identification of all types of unknown words

Key words: unknown word; Chinese segmentation; hybrid strategy; global statistic