局部统计在汉语未登录词辨识中应用和实现方法[1]

沈达阳* 孙茂松 黄昌宁

(*汕头大学计算机科学研究所，汕头515063 Email:dyshen@mailserv.stu.edu.cn)

(清华大学计算机系，北京100084)

摘要：真实文本中的未登录词辨识是中文信息处理中的新问题。本文讨论了文本局部统计在汉语未登录词辨识中的应用，探讨了局部缓冲大小和未登录词辨识性能之间的关系，并给出了一种局部统计的实现方法。实验表明，对于未登录词比较密集的真实文本，局部统计可以辨识出一半以上的未登录词。系统作为清华大学分词和词性标注系统SegTag的一个子系统，处理了一千五百万字的新华社通讯稿，显著提高了原来系统的切分精度。

关键词：中文信息处理 真实文本 未登录词辨识 上下文 局部统计 局部缓冲

The application and implementation of local statistics in Chinese unknown word identification

Shen Dayang, Institute of Computer Science Shantou Univ. Shantou 515063

Email:dyshen@mailserv.stu.edu.cn

Sun Maosong, Huang Changning, Dept. of Computer Science Tsinghua Univ. Beijing 100084

Abstract: Unknown word identification in running text is a new problem in Chinese NLP. In this paper, we discuss the application of local statistics in unknown word identification, then we investigate the relationship between the size of cache and its performance and present an implementation approach. The experiment shows that, local statistics are able to identify more than half of the unknown words in running Chinese text. The system, as a subsystem of the word segmentaion and part of speech tagging integrated system of Tsinghua University AI lab, has successfully processed the Xinhua News Corpus , which contains more than 1500, 0000 Chinese characters, and improved the word segmentation accuracy dramatically.

Key words: Chinese NLP, running text, Unknown word identification, context, local statistics, cache, word segmentation and part of speech tagging.

1. 引言

汉语未登录词的辨识是目前汉语自动分词研究的主要问题之一。由于未登录词（即词典中没有收录的词）问题是开放的系统所特有的。随着自然语言系统开始面向真实文本，未登录词的辨识问题开始引起关注。汉语词汇是一个开放的集合，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。因此，未登录词的自动辨识是十分必要的。

针对某种特定类型未登录词的辨识，国内外同行和我们已经作了不少研究[1][2][3]。主要方法是：从语料库中抽取一些有用信息，如汉字在某类语料库中的频度、二元语法、

[1] 国家自然科学科学基金重点项目资助。合同号：69433010

用词规律等，然后再总结相应的辨识算法。但是，这些研究都没有涉及到真实文本上下文信息的利用。

语言具有局部的统计特征，在真实文本局部的范围内，某些用词可能频繁地出现。在新闻语料中，许多未登录词，如人名，地名，如果和新闻的内容有关，其出现频率有时甚至超过常用词。因此，统计局部文本中字符串的出现频率，对未登录词的辨识十分有用。本文着重讨论了局部统计在汉语未登录词辨识中的应用，探讨了局部缓冲大小和未登录词辨识性能之间的关系，并给出了一种局部统计的实现方法。

2.局部频率在未登录词辨识中的应用

所谓局部频率（**local frequency**），就是局部文本范围内，某字符串的出现频率。为了统计字符串的局部频率，需要把一定大小的文本放在内存中；称为局部缓冲**(cache)**。局部频率可以在下面几个方面发挥作用。

2.1.单独利用局部频率来辨识未登录词

3.

局部缓冲	说明
<以色列一部长被革职>　　据突尼斯非洲通讯社报道，以色列总理沙米尔今天宣布，解除科学和研究部长埃采尔·韦茨曼的职务，理由是“他最近同巴解组织进行了接触”。这项决定是在以色列内阁举行每周一次的会议后于今天上午宣布的。　　韦茨曼现年65岁，以色列工党成员，70年代末曾参加过埃及－以色列和谈。他经常对以色列当局提出批评，主张以色列同巴解组织谈判。	在真实文本中，可以经常发现未登录词频繁的出现在局部的文本范围内，如在左边的一小段话中(11个句子)，未登录词"以色列"就出现了7次.

因此，寻找真实文本局部范围内的高频字符串，是一种发现未登录词的有效方法。其步骤如下：

①从文本中读取一定数目的句子，放在内存中，称为局部缓冲;

②对于每个句子，经过粗分词之后，在分词碎片中，搜索出频率大于2的字符串;

③有时会出现冲突现象，在这种情况下，先选取频率高者，若还有冲突，再选取长度长者，再有冲突，就都保留下来。这样搜索出来的高频字符串就是利用局部统计猜测出来的未登录词（以后称为局部猜测）。下面是一个例子：

例句（来自上面的局部缓冲中）	局部频率
这项决定是在以色列内阁举行每周一次的会议后于今天上午宣布的.	f(在以)=2，f(以色列)=7，f(以色)=7，f(色列)=7， 局部猜测取:以色列

2.2. 和其他特定类型未登录词辨识方法相结合

在中国地名，中国人名，外国人名识别[1][2][3]中，未登录词边界的确定标准可以根据局部频率的大小作相应的改变，从而可以使辨识的精确率进一步提高。如：

中国地名的识别	
句子	神府东胜煤田地跨陕西省和内蒙古自治区
切分结果	神/府/东/胜/煤田/地/跨/陕/西/省/和/内/蒙/古/自治区
原来的辨识结果	神府，东胜，陕西省，蒙古自治区

错误原因	“内”在中国地名库中出现的次数很少
局部频率的运用	若"内蒙古"的局部频率较高，则不排斥"内"

外文译名的识别	
句子	9号纽厄尔门前从容射入一球
切分结果	9号/纽/厄/尔/门/前/从容/射/入/一/球
原来的辨识结果	纽厄尔门
错误原因	“门”在外文译名库中出现的次数较高
局部频率的运用	若"纽厄尔"向前搜索到"纽厄尔门"时，局部频率下降，则不再往前搜索

中国人名的识别	
句子	李鹏飞抵香港
切分结果	李/鹏/飞/抵/香/港
原来的辨识结果	李鹏飞
错误原因	“飞”在中国人名库中出现的次数较高
局部频率的运用	若"李鹏"向前搜索到"李鹏飞"时，局部频率下降，则不再往前搜索

2.3.几个未登录词连续出现时，提供正确切分的依据

在局部缓冲中，有下面的普遍规律：

局部缓冲中有一字符串W，W被某种特定类型的未登录词辨识方法作为未登录词识别出来。W=AB，A、B是W的两个子串，其局部频率分别为F(A)、F(B)，若F(A)≠F(B)，那么W可以切分为A、B两个未登录词。因为:

若F(A)≠F(B)，必有F(A)>F(B)或F(A)<F(B)，也必有F(A)≥2或F(B)≥2;那么，

①在局部缓冲中，必有A或B单独出现的情况，而且出现了两次以上，一般地说，A或B就是一个词;

②由于W被作为未登录词(中国地名，中国人名或外文译名)识别出来，说明A和B同时还具备 了该种类型未登录词的数据分布特点。

从上面两点看，A和B是未登录词的可能性是很大的。当然，单纯从语言的角度看，或许可能举出反例来，但在真实文本中，我们尚没有发现这个规律的反例。具体如：

例文	说明
据外电报道，第十四届世界青年冰球锦标赛第四轮比赛１２月３１日在芬兰赫尔辛基结束，苏联队和捷克斯洛伐克队积分领先。　　在３１日举行的两场比赛中，瑞典队以１４：０大胜波兰 队。前三轮与苏、捷两队同处领先地位的加拿大队和东道主芬兰队以３：３战平，积分落到了苏、捷两队之后。	句子“..在芬兰赫尔辛基结束...”中，“芬兰赫尔辛基”作为一个外文译名被辨识出来，由于下文中有“...东道主芬兰队以...”，根据上面的规律: W=芬兰赫尔辛基， A=芬兰，F(A)=2 B=赫尔辛基，F(B)=1 所以A，B是两个未登录词

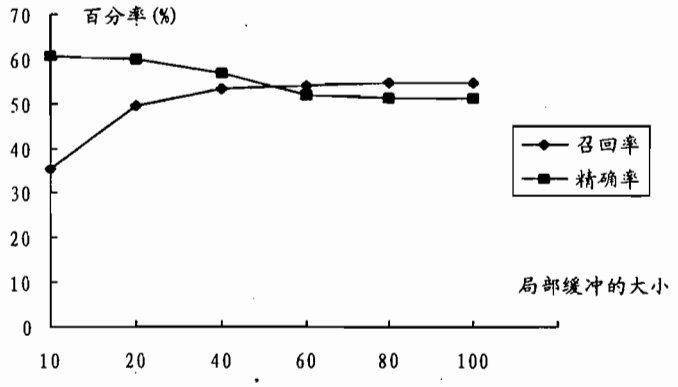
<宁夏艺术团在尼首场演出> 中国宁夏回族自治区艺术团１２月３１日晚在这里进 行了首场演出。 尼泊尔首相什雷斯塔观看了演出并向艺术团赠送了花篮。今天，艺术团为尼泊尔观众表演了杂技、歌曲及民族舞蹈等节目。演员们精湛的技艺博得１０００多名观众的阵阵掌声。 宁夏艺术团应尼泊尔皇家文学院的邀请于昨天抵达这里，进行访问演出。	句子“中国宁夏回族自治区艺术团...”中，“中国宁夏回族自治区”作为一个地名被辨识出来，由于上下文中各出现“宁夏”一次，根据上面的规律： W=中国宁夏回族自治区 A=中国，F(A)=1 B=宁夏，F(B)=3 C=回族自治区，F(C)=1 所以，A，B，C是三个未登录词

3. 局部缓冲大小和未登录词辨识性能之间的关系

为了考察缓冲大小和猜测结果的关系，我们对一个长度为12k的新闻文本作了实验。下面是实验的结果:

未登录词总共:204个				
缓冲的大小(句子数)	猜测的未登录词数	正确的未登录词数	召回率(%)	精确率(%)
10	119	72	35.3	60.5
20	169	101	49.5	59.8
40	192	109	53.4	56.8
60	212	110	53.9	51.9
80	218	112	54.9	51.4
100	218	112	54.9	51.4

下图表明缓冲大小和猜测召回率，精确率的关系



由上图可见，当局部缓冲的大小在>80时，召回率和精确率已渐趋稳定。虽然，局部缓冲的大小在>80时，精确率有所降低，但由于我们有后续的手段可以进一步处理，相比之下，召回率更重要些，所以，实验中我们把局部缓冲的大小定为100。另外，就新闻语料而言，一半以上的未登录词（54.9%）可以靠局部统计辨识出来。

在文献[4]中，通过一个评价函数，把文本划分为统计意义上的段落，没有上述决定局部缓冲大小的问题。但其中的数据都要求在读入所有文本的前提下获得的，所以，测试文

本都有长度的限制。另外，其寻找未登录词“岛屿”的算法复杂度也是极高的，不适合在分词系统中使用，因此，本文没有采用该方法。

4. 局部猜测的后处理

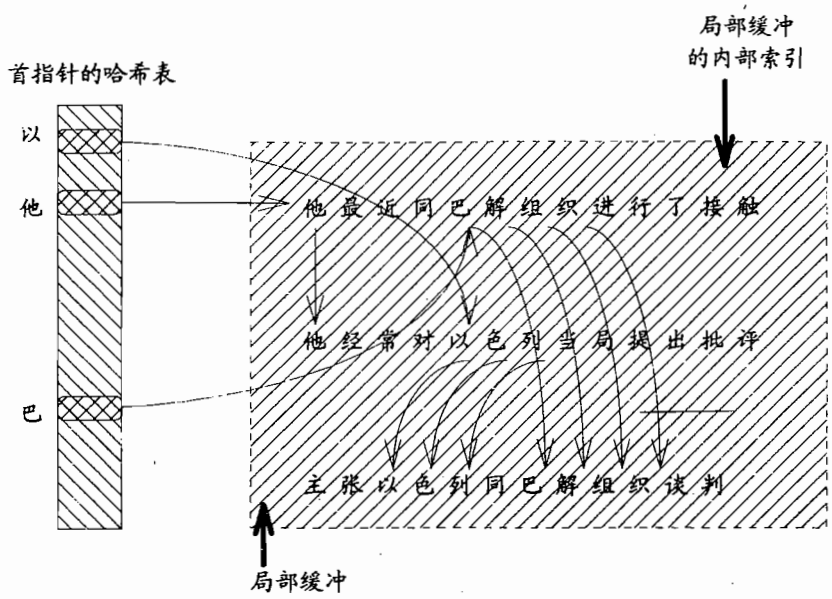
错误的局部猜测主要有下面两种情形：①含有若干虚词的字段；②有几个高频单字词组成的字段；

含有虚词	（小标题）防止藻类附着船底的新技术　日本三菱重工业公司最近发明了防止海洋中藻类和贝类等附着船底的新技术。...	局部猜测:的新
	从而为研究黄曲霉毒素与肝癌的 关系提供了一个必要的手段。....早在１９８９年的第六届全运会卜．该电子信息中心首先开发了一套电脑管理系统。	局部猜测:了一
连续几个高频单字词	政协全国委员会今天上午在政协礼堂举行新年茶话会。　党和国家领导人江泽民、李鹏、万里、姚依林、宋平、李瑞环、王震等同全国政协、各民主党派、无党派爱国人士和全国工商联负责人以 及各界人士欢聚一堂，共贺新年。	局部猜测:全国
	中共中央总书记江泽民在茶话会上说，今年是治理整顿、深化改革的 关键一年，又是九十年代的开始。做好今年的工作对我国今后的发展具有十分重要的意义。第一位的是要保持社会的稳定。 社会稳定了我们才能集中精力做好各项工作。	局部猜测:做好

这些错误可以用规则和全局统计（互信息，T－测试等）来清除[5][6]，清除了上面两种错误之后，在召回率基本不变的情况下，局部猜测的正确率可以提高到88.2%。剩下的错误主要是数量词的组合，如“第二轮”，“第一届”等，在分词系统中，可以通过预切分进一步清除[6]。

5. 局部缓冲的实现方法

由于局部缓冲不仅仅用于搜索局部文本中的高频未登录词，它在其它未登录词的辨识中也有一定的辅助作用，因此，局部缓冲的搜索速度对于分词系统的速度十分关键。如果简单地把自由文本放在内存中作盲目的搜索，对系统的整体切分速度影响很大 ，尤其是当局部缓冲的较大时，对系统性能的影响是难以忍受的。因此，具体实现时，我们采用了下面的一些加速措施。



说明： 局部缓冲除保留局部文本的空间外，还有与该缓冲相对应的内部索引空间，以及一张哈希表。当系统读入文本之后，把哈希表中的指针指向读入文本中与该指针相应的第一个汉字，当需要在局部缓冲中搜索一个字符串的出现次数时，经过一次哈希运算就可以找到该字可能出现的第一个位置，再利用内部索引陆续搜索可能的其它位置。这种搜索方法的速度是盲目搜索的S/N倍，S是缓冲的大小，N是字符串可能 出现的次数。当缓冲大小为100个句子时若平均句长为10，S≈100 • 10，N≈1，速度可以提高1000倍左右。

6. 结束语

本文讨论了文本局部统计在汉语未登录词辨识中的应用，并实现了一个基本局部统计的未登录词辨识系统，实验表明，对于未登录词比较密集的真实文本，局部统计可以辨识出一半以上的未登录词。本文还进一步探讨了局部缓冲大小和未登录词辨识性能之间的关系，并给出了相应的实验数据。该系统作为清华大学分词和词性标注系统SegTag的一个子系统，处理了一千五百万字的新华社通讯稿，显著提高了原来系统的切分精度。

【参考文献】

[1] 孙茂松、张维杰，英文译名的自动辨识，《计算语言学研究与应用》，北京语言学院出版社1993.

[2] 孙茂松、黄昌宁、高海燕、方捷，中文姓名的自动辨识，《中文信息学报》，第9卷，第2期，1995.

[3] 沈达阳、孙茂松、黄昌宁，中国地名的自动辨识，《计算语言学进展和应用》，清华大学出版社1995.

[4] 白栓虎，汉语词切分及词性自动标注一体化方法，《计算语言学进展和应用》，清华大学出版社1995.

[5] 沈达阳，基于语料库的汉语未登录词发现，学士学位论文，清华大学，1993.

[6] 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和实现，硕士学位论文，清华大学，1996.