面向中文陌生文本的人机交互式分词方法

李斌，陈小荷

（南京师范大学文学院，南京 210097）

摘 要： 本文提出了面向中文陌生文本的人机交互式分词方法，在没有分词底表和训练语料等语言知识的条件下，由系统自动地发现未登录词，提交给用户进行增删，不断重复此过程，反复获取文本中的词语，最后进行最大匹配法分词。四个不同语料的实验显示，在没有人机交互的条件下，可以得到72%（F值）左右的分词精度。经过较少的人机交互，可以使分词F值提高12%以上。随着用户工作量的增加，系统还能够进一步提高分词效果。

关键词： 陌生文本；人机交互；自动分词；未登录词识别；中文信息处理

A HCI Word Segmentation Method Adapting to Chinese Unknown Texts

LI Bin, CHEN Xiaohe

(School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097)

Abstract: This paper proposed a novel method requiring no lexicon or hand-crafted linguistic resource. It can deal with various texts and can adapt to different WS standards. Through Human-Computer Interaction (HCI), candidates as word recursively extracted from the text, are judged and edited by the user. Thus, a lexicon of the text is gained and then applied to segment the text. Experiments on 4 different texts show that without HCI, F-score of our system reaches as much as 72%, and can be prompted by 12% with amount of work done by the user. With the increase in the workload of the user, the system is able to achieve better results.

key words: unknown text; Human-Computer Interaction; word segmentation; unknown word recognition; Chinese Information Processing

1 引言

自动分词是中文信息处理的基础课题之一。随着中文电子文本数量的日益增加，文本的领域呈多样性发展，语料库的加工要求也有所不同。Zhongjian Wang et al.（2002）指出，一个分词系统也应当能够处理不同领域的文本和适应不同的分词标准。对服务于汉语研究的语料库加工而言，如何对现有的大量古代汉语的电子文献进行分词，如何对珍贵的方言语料进行处理等等，都是亟需解决的问题。在此背景下，本文提出了面对中文陌生文本的人机交互式分词方法。所谓“陌生文本”，即对于分词系统来说，没有关于该文本的任何词汇、句法、语义等先验的语言知识和资源。所谓“人机交互”，就是由系统自动地从文本中获取候选字串，由用户根据其上下文进行

基金资助：南京师范大学211资助项目 语言信息处理与分领域语言研究的现代化（1240702504）

作者简介：李斌（1981-），男，江苏徐州人，博士研究生，E-mail：gothere@126.com.

筛选，得到适应于不同领域的词语特点和分词标准的词表。面向陌生文本的分词，就是让系统在没有词表和其他资源的条件下，通过人机交互的方式完成对汉语各种文本分词处理。

2 相关工作

目前，作为主流的基于统计分词的方法所关注的是如何从训练语料中尽可能多地学习语言知识，再对其他的同质文本（“非陌生”文本）进行分词。因此，该类无法适用于陌生文本的自动分词。而不需要词表和训练语料等资源的陌生文本分词技术研究较少，还处在实验阶段。王开铸 等（1995）使用统计方法从待切分语料中抽词，又将所抽取的词条用于自动分词。黄萱菁 等（1996）利用 x^2 统计量进行自动分词。傅赛香 等（2002）使用了串频统计方法，然后通过长短串的频次的比值进行过滤获得词表，再进行分词。Xiaopeng Tao et al.（2003）则建立了一个文本熵的模型，其原则是文本分词的结果越好，则文本的整体熵越低。这些方法是纯粹利用统计方法进行陌生文本分词的一个尝试，分词的精度既不高也不够稳定。因此，一些学者考虑使用人机交互的方式来增加系统的语言知识。Sun Maosong et al.（1995）利用邻接汉字的统计信息，让机器自动地给出针对该语料的候选词表，然后由用户进行筛选。通过迁值控制，以半自动循环的工作方式，最终得到一个词表。该文没有进一步进行全文分词，但其提出的人机交互式的方法，可以保证获取词表的精确率，缺点是召回率难以保证。

较为实用的陌生文本分词方法则是 Zhongjian Wang et al.（2002）提出的基于句子的人机交互的增量式学习方法。首先，使用串频统计，获取文本中的未登录词，然后，利用这个词表进行自动分词，再把分词结果提交人工判定，利用学习到的词语和优化参数进行下一轮分词和未登录词的提取。在规模为 9 万词的语料上，可以达到近 90%的分词正确率。然而，其未登录词的发现性能较差，在人工判定的条件下，只能达到 30%左右的正确率和召回率，大量的工作实际上还是通过人工判定来完成。冯冲 等（2006）提出了基于 Multigram 语言模型的主动学习分词方法，也是基于句子的学习，依靠对较为高频的句子和词语进行学习，解决高频字串的切分问题。

总的来看，人机交互的方式比纯统计方法的效果要好。让用户来确定词，可以让系统适应于不同的分词标准。然而，这些方法存在的最大问题就是未登录词发现的精确率和召回率不高，在人机交互和机器自动学习的机制上存在一些问题，导致最后的分词效果不好或代价过高。

3 算法

本文采用基于候选词方式进行人机交互。上文介绍了人机交互的两种交互方式，基于句子的和基于候选词的，这两种方式各有其优缺点。前者可以得到切分好的句子集合，但对于用户而言，切分整个句子比较困难一些。相当数量的词会反复出现在不同的句子中，造成人工的浪费，也容易出现对同一个词切分不一致。同时，要定义生成候选句子的判别函数也是比较困难的。基于候选词的交互方式则可以直接得到该语料的词表，而且通过观察上下文，能够让用户比较容易判定是否是词，同时也可以避免用户对同一个词的切分不一致。因此，我们采用了基于候选词的交互方式进行自动分词。

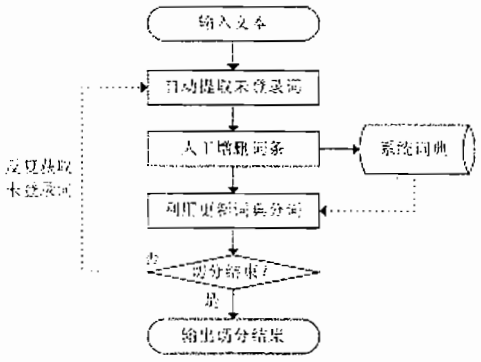

图 1. 系统流程图

3.1 系统流程

图 1 给出了系统流程。首先，由机器从陌生文本中自动抽取一个高精度的词表。接着，由熟悉该文本的专业

人员或用户进行词条的甄选，得到一个小规模词表。然后，利用这个词表进行自动分词，在未切分的汉字串中，抽取出更多的词条，由人工进行判定。不断地重复这个人机交互的过程，最终完成对文本的分词。

使用统计方法进行自动抽词，必然面对两个问题，即探测未登录词
和适应不同的分词标准。从技术上考虑，这两个问题可以看作两个过程，一是如何自动地提取候选字串并进行自动筛选，保证词语的精确率和召回率；二是如何通过人机交互来确定词表，让词语符合用户的分词标准。本文提出了改进的后缀数组抽词算法，对抽取的候选词语采用互信息（MI）进行过滤，得到了性能较好的自动抽词模块。同时，提供较好的人机交互界面，便于用户增删词语。

3.2 改进的后缀数组自动抽词算法

提取候选字串时，最大的问题是会产生大量垃圾。如，假设在一个文本中，字串“萨达姆”、“萨达”和“达姆”的频次都为 10 次。很明显，“萨达”和“达姆”是需要过滤的字串。针对这一问题，目前主要有两种做法：一种是基于哈希表，计算文本中所有的 n 元字串的频次，然后使用频次相减法来过滤（金翔宇 等 2001）。另一种是利用排序的后缀数组，直接把数组序列中前缀相同的字串提取出来，这样可以排除掉“萨达”。同样地，建立排序的前缀数组来排除“达姆”（Luo Zhiyong et al. 2004）。这两种方法，在时间和空间上开销过大，难以满足实用系统的需要，也无法提取频次为 1 的词语。为了解决计算效率问题，我们提出了改进算法，只需建立一个排序的后缀数组，就可以完成排除子串的过程。首先，利用排序的后缀数组，可以排除“萨达”。然后，进一步利用上文的后缀信息来提取候选串。如果上文有相同的字符，则不算作候选字串。由于“达姆”所在的后缀数组，其上文必定为“萨”，可以被排除。通过计算它们上文相同的长度，即上文最长公共后缀（LLCS，Longest Left Common Suffix）的长度，就可以跳过这些被长串完全覆盖的子串。在不增加空间开销的前提下，把算法的时间复杂度由原来的$O(N^2)$降到了$O(N*\lg N)$。由此，利用邻串的 LCP、LLCS 值，可以从文本中自动获得大量的 n 元字串。图 2 是从 1998 年 1 月人民日报语料中提取出来的一部分以“中国”开头的后缀数组。

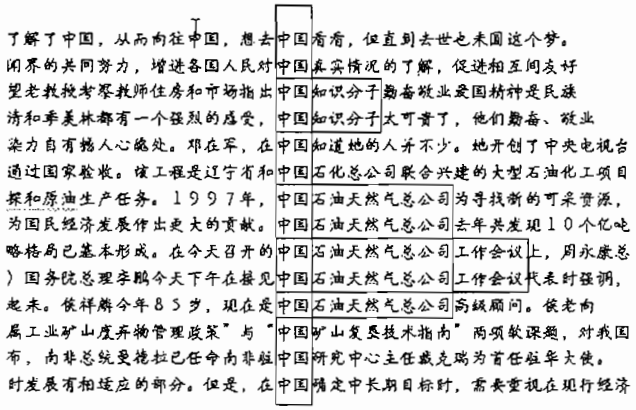

图 2 1998 年 1 月语料中“中国”的排序后缀数组示例

利用改进的后缀数组提取 n 元字串，只能解决长串包含短串的问题和系统的时空开销问题，并不能直接提取出真正的词，依然存在以下问题：

1） 只能提取频次为 2 以上的 n 元字串，导致频次为 1 的词无法提取。

2） 跳跃出现相同的字串，频次需要累加。如，上图中的多处出现的“中国”。

3） n 元字串大于真正的词，边界不好确定，这种情况数量庞大，需要筛选和过滤。如，“中国石油天然气总公司工作会议”。

4） n 元字串小于真正的词，边界不好确定，这种情况数量很少，如“哥伦比亚”、“无与伦比”造成的“伦比”。

针对这四个问题，我们提出了相应的解决方案：

1） 使用左右扩展法在后缀数组中提取频次为 1 的低频字串。方法是利用频次为 1 的二字串进行左右扩展。对汉字串 ABCDE，如“起诉萨达姆”中，假设“起诉、诉萨、萨达、达姆”的频次分别为 5、1、1、1。假设以“萨达”为出发点，即 CD 的频次为 1，向左扩展，如果 BC 的频次为 1，扩展为 BCD（诉萨达）；再往左扩展，如果 AB 的频次大于 1，则删除 B；以此类推不断向左扩展，确定左边界；以相同的方式可以确定右边界。同时屏蔽掉后缀数组中其他出发点的二字串。最后得到频次为 1 的“萨达姆”。需要说明的是，该方法也只能获取一

部分频次为 1 的字串。

2） 采用哈希词典的方式存储候选字串，把相同字串的频次累加。

3） 过滤长串。互信息经常用于衡量相邻字符结合的紧密程度。互信息越高，则成词的可能性越大，可以用来判定一个字串是否为词。利用多元互信息公式计算候选串的 MI 值，低于设定的阈值则删除。

4） 过滤短串。由于这种情况出现的不多，可以直接提交给用户进行处理。

3.3 互信息过滤

对于 2 元至 4 元字串采用条件熵推导出来的互信息公式，对于 5 元以上的字串，由于公式过于繁琐，计算量过大，我们采用另一个简化公式。计算公式如下，

$$MI(a;b)=\log\frac{P(ab)}{P(a)P(b)} \quad (n=2)$$

$$MI(a;b;c)=\log\frac{P(ab)P(bc)P(ac)}{P(a)P(b)P(c)P(abc)} \quad (n=3)$$

$$MI(a;b;c;d)=\log\frac{P(ab)P(bc)P(ac)P(ad)P(bd)P(cd)P(abcd)}{P(a)P(b)P(c)P(d)P(abc)P(abd)P(acd)P(bcd)} \quad (n=4)$$

$$MI(C_1;C_2;...;C_n)=\log\frac{P(C_1,C_2,...,C_n)}{(P(C_1)*P(C_2)*\cdots*P(C_n))^{\frac{1}{n-1}}} \quad (n>4)$$

其中，$n\geq 1$，$f(C_1,...,C_n)$ 是 n 元字串 $C_1,...,C_n$ 在语料中出现的次数，N 是语料规模（总字数），$P(C_1,...,C_n)=\frac{f(C_1,...,C_n)}{N}$。

$P(ac)$，则是字符 a 和 c 相隔一个字符时顺序共现的概率。

为了提高自动抽词的性能和效率，我们在系统中加入了识别两种特殊字串的独立模块。一种是由汉字构成的“AABB”型重叠式，如“风风火火”等词语。一种是“简单数词”，包括阿拉伯数字、汉字数字构成的数词。如，“3000”、“20 万”、“叁拾”等。这二种字串在汉语文本中经常出现，可以作为未登录词识别模块的补充，用于人工筛选。

3.4 人机交互

人机交互是让用户借助上下文信息判定一个候选字串是不是词，以得到质量较高的词表，进行后续的抽词和分词流程。我们规定，用户在筛选候选词时，只能进行三种操作，即“确定”、“删除”和“添加”。如果候选串是词，则进行“确定”操作；不是词，则“删除”；在上下文观察时发现新的词，则“添加”到词库中。

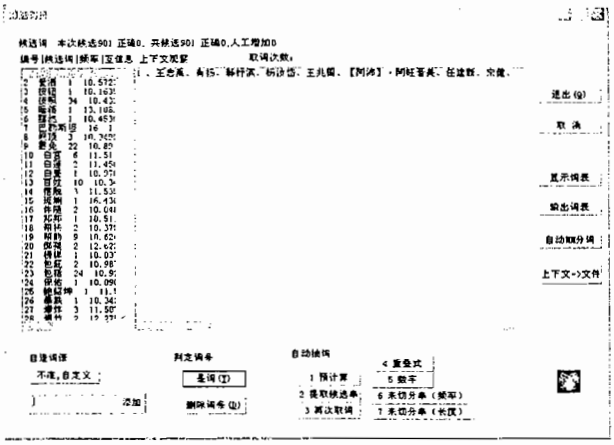

图 3 人机交互式分词界面

我们在 VC6.0 环境下实现了该系统。人机交互界面如图 3，在 pku_test（SIGHAN2005）上，系统第一次自动抽词得到的词语列表，按音序排列，并给出频次和互信息值。单击“阿沛”，右侧则显示出其上下文，用户可以根据自己的要求进行增删词条。

4 实验结果及分析

我们对四种规模、体裁、分词标准各不相同的语料进行测试，其中包括普通分词系统难以处理的现代汉语和近代汉语的小说语料。

测试语料

语料	繁简体	名称	语料规模	体裁	分词标准	来源
现代汉语语料	简体	pku_test	149886 字	新闻	北大	SIGHAN2005[3]
	简体	英雄出世	54594 字	小说	北大	南师大[4]
	繁体	as_test	172181 字	新闻	台湾	SIGHAN2005
近代汉语语料	繁体	红楼梦	718388 字	小说	台湾	台湾中央研究院[5]

4.1 测试方法

为了说明人机交互的效果，同时避免人的主观操作的不稳定性，我们采用与答案词表（即从分词语料中提取出来的词表）进行比对的方式，模拟用户的操作过程。系统模拟用户自动地进行“确定”和“删除”操作时，只需通过查询答案词表即可实现，而对于“添加”操作，系统只能在“候选串”的内部和外部上下文中进行未登录词的查找。

内部查找：对于一个候选字串 S，在其内部查找所有的未登录词，收入“已知词表”。

外部查找：在 S 所在的 100 条上下文中寻找五种简单模式的词语添加到“确定词表”中，其它情况的未登录词则不再收入“已知词表”。C_{-2}、C_{-1} 为 S 的上文 2 个字，C_1、C_2 为下文的 2 个字，5 条横线即为查找未登录词的 5 种模式。

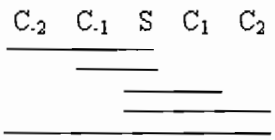

图 4 字串的外部查找模式

4.2 评测标准

我们从未登录词识别效果、分词精度、用户劳动量等三个方面进行评测。未登录词和分词的正确率、召回率、F 值比较容易测算。而对于用户劳动量，我们以用户在进行交互时花费的总时间为依据，采用了加权的方法进行计算。在用户对语料比较熟悉和软件操作熟练的情况下，“确定”、“删除”和“添加”三种操作的平均时间的经验值约为 1 秒、2 秒和 3 秒。因此，工作时间=“确定”条数*1+“删除”条数*2+“添加”条数*3。

为了突出多字词的抽取效果。表 2 中候选串的正确率、召回率、F 值的计算都是以答案词表中的多字词条数作为分母。此外，我们还给出了用于比较分词性能的 Baseline 和 Topline。Baseline 是没有经过用户筛选，系统反复获取未登录词，而后进行正向最大匹配法（FMM）分词得到的 F 值。Topline 则是使用答案词表进行 FMM 分词得到的 F 值。

4.3 测试结果及分析

表 2　　四个语料上的测试结果

语料	pku_test	英雄出世	as_test	红楼梦
总词数（条）	13148	5007	18812	19003
多字词（条）	11672	3738	17246	15921

[3] SIGHAN2005 分词语料的下载网址：http://www.sighan.org/bakeoff2005/。

[4] 《英雄出世》的分词语料由南京师范大学计算语言学专业 2004 级硕士加工。

[5] 《红楼梦》的分词标记文本惠蒙台湾中央研究院友情提供。该语料的特殊问题：在中央研究院加工该语料时，出于研究近代白话文的目标，已将其中的诗词部分全部删除。原文为 Big5 格式，中研院在加工时有 100 多个汉字无法录入，文本中采用●等符号代替。在切分语料中，每一回的“回目”，即标题没有切分，数词也全部切为单字。如：“十 六 歲”。

交互情况	候选串（条）	5926	3376	9089	15713
	候选正确（条）	3797	1684	6824	6374
	正确率	0.6407	0.4988	0.7508	0.4057
	召回率	0.3253	0.4505	0.3957	0.4004
	F 值	0.4315	0.4734	0.5182	0.4030
	用户添加（条）	1599	399	2770	1968
	得到总词表（条）	5396	2083	9594	8342
	总召回率	0.4623	0.5572	0.5563	0.5240
	交互次数	19	18	22	22
	工时估算	3.6 小时	1.7 小时	5.5 小时	8.6 小时
分词精度	正确率	0.8576	0.8469	0.8019	0.9354
	召回率	0.9210	0.8644	0.8977	0.9512
	F 值	0.8882	0.8556	0.8471	0.9432
分词 Baseline	未经人机交互	0.7606	0.6759	0.7217	0.7226
分词 TopLine	答案词表+FMM	0.9752	0.9120	0.9664	0.9738

四个语料的测试结果显示，未经人机交互时，系统分词的 F 值已经可以达到 70%左右。在较少的人工耗费下，分词 F 值可以达到 84%以上。其中，《红楼梦》语料，由于单字词出现的比例较大，得到的分词 F 值最高（94%）。相对于 Baseline，在花费了一定的人力进行交互以后，系统的分词性能确有不小的提升，F 值分别提高了 12 个百分点以上。而更为重要的是，用户通过筛选获得了一个符合自己的分词标准的相当规模的领域词表。当然，相对于 Topline 来说，人工交互的分词效果还有待进一步提高。

从未登录词的获取情况来看，不同的语料差别较大，尤其是正确率、召回率和用户添加词语的比例相差很多。但是仔细分析后可以发现，抽取的候选串的 F 值基本相近，系统得到的词（候选正确+用户添加）的总召回率，也都保持在 50%的水平上，这可能是由于语料的特殊性造成的。同时，文本中依然有 50%左右的多字词没有识别出来，主要是低频词语，说明系统在获取低频词语方面还需要改进。

由于在模拟人机交互时严格限定了“添加”词条操作的范围，使得人机交互的最后结果不够理想。在实际操作时，系统还允许用户使用其他先验的词表，或者自行添加一部分词语，从而得到更好的分词精度。

为了进一步提高系统性能，我们还专门设计了用于提取未切分串中重要信息的模块。使得分词精度随着用户干预的增加而不断提高。经过多次人机交互后，达到互信息的最低阈值，会导致无法继续提取候选串的情况。此时，如果把未切分串中的高频条目进行人工判定，则可以从中提取出系统词表没有收录的中高频未登录词，提高系统性能。进一步地，把未切分串中长度大的条目提交给用户判定，可以得到中低频的未登录词，丰富系统词表。使用这两个模块后，系统的分词性能会有提升。由于使用该模块并不能直接得到未登录词，而是一些长度较大的字串，几乎完全依靠用户来判定和添加到词表中，因此，该模块仅作为用户选用的一项辅助措施，没有参与评测。

5 结论与未来工作

本文提出了面向中文陌生文本的分词方法，在没有分词底表、训练语料和其他语言知识的条件下，可以根据用户的标准进行分词。该方法采用人机交互的方式，不断扩大系统词表，尽可能地获取文本中的所有词语，从而达到较高的分词精度。系统以 Unicode 字符集为核心，可以处理不同编码（繁简体）的文本。在文本的通用性上，可以处理不同时代（现代汉语、近代汉语）、不同领域（新闻、文学等）的汉语文本，从而为特殊语料库的加工提供了一个较为高效的分词工具。在未登录词发现方面，重点解决了使用后缀数组时排除长短串覆盖问题和频次为 1 的字串提取问题。

文本是对陌生文本进行分词的一次初步尝试，还存在一些不足和需要进一步研究的问题。如，提高低频词语的识别效果；探索更好的人机交互方式；增强系统的智能性，更好地利用用户反馈的信息，减少用户的工作量；解决文本中存在的切分歧义；进一步开发出人机交互式的词性标注系统、义项标注系统，从而使古代汉语文本和

汉语的其他特殊文本的深加工能够在一个较为高效和智能的平台上展开。

致谢

感谢微软亚洲研究院的黄昌宁教授、北京工业大学的宋柔教授、清华大学的孙茂松教授在论文思路上的指导，台湾中央研究院的黄居仁、魏培泉教授为本文提供的语料，以及各位同门师兄弟的热心帮助。

参考文献：

[1] Zhongjian WANG，Kenji ARAKI，Koji TOCHINAI. A Word Segmentation Method with Dynamic Adapting to Text Using Inductive Learning[A]. In：Proceedings of the First SIGHAN Workshop on Chinese Language Processing[C]，2002：113-117.

[2] 王开铸，李俊杰，吴岩. 无词典自动分词的研究[A]. 陈力为，袁琦主编. 计算语言学进展与应用[C]. 北京：清华大学出版社，1995.

[3] 黄萱菁，吴立德，王文欣，等. 基于机器学习的无需人工编制词典的切词系统[J]. 模式识别与人工智能. Vol. 9，No. 4，1996：297-303.

[4] 傅赛香，袁鼎荣，黄伯雄，等. 基于统计的无词典分词方法[J]. 广西科学院学报，Vol. 18，No. 4，2002：252-255.

[5] Xiaopeng Tao，Shuigeng Zhou. Chinese Word Segmentation Without Auxiliary Data[A]. Maosong Sun，Tianshun Yao，Chunfa Yuan. In：Advances in Computation of Oriental Languages [C]. Beijing：Tsinghua University Press，2003：88-94.

[6] Sun Maosong，Shen DaYang.，Hang Changning. Deriving Chinese Lexicons from Large Corpora[A]. In：NLPRS-95，Tacjon，Korea，1995.

[7] 冯冲，陈肇雄，黄河燕，等. 基于 Multigram 语言模型的主动学习中文分词[J]. 中文信息学报，Vol. 20，No. 1，2006：50-58.

[8] 金翔羽，孙正兴，张福炎. 一种中文文档的非受限无词典抽词方法[J]. 中文信息学报，Vol. 15，No. 6，2001：33-39.

[9] Luo Zhiyong，Song Rou. An Integrated Method for Chinese Unknown Word Extraction[A]. In：Proceedings of 3rd ACL SIGHAN Workshop [C]. Barcelona，Spain，2004：148-154.