文章编号: 1003-0077(2007)03-0092-07

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,江苏 南京 210097)

摘 要: 自动分词是中文信息处理的基础课题之一。为了克服传统分词方法在处理特殊领域文本时遇到的困难,本文提出了一种新的分词方法,在没有词表和训练语料的条件下,让用户参与到分词过程中,增加系统的语言知识,以适应于不同的语料和分词标准。系统采用改进的后缀数组算法,不断提取出候选词语,交给用户进行筛选,最后得到词表进行分词。四个不同语料的实验结果显示,不经过人工筛选,分词 F 值可以达到 72 %左右;而经过较少的人机交互,分词 F 值可以提高 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: 计算机应用;中文信息处理;自动分词;未登录词识别;陌生文本;人机交互

中图分类号: TP391 **文献标识码**: A

A Human-Computuer Interaction Word Segmentation Method Adapting to Chinese Unknown Texts

LI Bin, CHEN Xiao-he

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

Abstract: Word segmentation(WS) is a funamental task in Chinese information processing. To solve the difficulties of traditional methods in processing texts in restricted domains, a novel method is proposed. It requires no lexicon or training corpus and can adapt to various texts and different WS standards. It enables the user to take part in WS procedure and add language kownledge to the system. Using optimized suffix array algrithm, candidates as words are recursively extracted from the text, then judged and edited by the user. Thus, a lexicon of the text is gained and applied to segment the text. Experiments on 4 different texts show that without the user's judgement, F-score of the 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: computer application; Chinese information processing; word segmentation; unknown word recognition; unknown text; Human-Computer Interaction

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。文献[1]指出,一个分词系统应当能够处理不同领域的文本和适应不同的分词标准。对于以汉语研究为目的的语料库建设而言,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的分词处理。

收稿日期: 2006-08-28 **定稿日期**: 2007-02-14

基金项目: 南京师范大学 211 资助项目(1240702504)

作者简介: 李斌(1981—),男,博士生,主要研究方向为计算语言学。

2　相关工作

目前,作为主流的基于统计的分词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,再对同质文本(“非陌生”文本)进行分词。因此,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。文献[2]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。文献[3]利用 χ^2 统计量进行自动分词。文献[4]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。文献[5]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。文献[6]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。通过阈值控制,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文献[1]提出的基于句子的人机交互的增量式学习方法。首先,利用串频统计获取文本中的未登录词,然后,基于这个词表进行自动分词,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不高,在人工判定的条件下,正确率和召回率分别为 26 %和 31 %,大量的工作实际上还是通过人工判定来完成。文献[7]提出了基于 Multigram 语言模型的主动学习方法,首先使用了 50M 同质生语料利用 EM 算法来参数估计,再依靠对较为重要的句子提交用户切分,解决高频字串的切分问题。在开放测试中,分词 F 值为 77.7 %。

总的来看,在处理陌生文本时,人机交互的方式比纯统计方法的效果好。让用户来确定词,不仅较为准确,还可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,导致分词效果不好或代价过高。

3　算法

上文介绍了人机交互的两种方式,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。而基于候选词的交互方式则可以直接得到该语料的词表,通过观察上下文,能够让用户比较容易判定是否是词,也可以避免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。

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候选词表。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,抽取出更多的候选串,由人工进行判定。这样反复进行人机交互,最终完成对文本的分词。

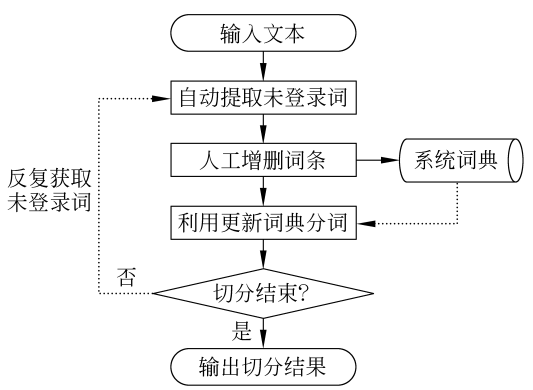

图 1　系统流程图

上述过程主要分为两个问题,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。例如在一个文本中,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”是需要排除的字串。针对这一问题,目前主要有两种做法:一种是计算文本中所有的 N 元字串的频次,然后使用频次相减法来过滤[8]。另一种是分别建立前缀数组和后缀数组。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,这样可以得到“萨达姆”,排除“萨达”。同样地,建立排序的前缀数组来把后缀相同的字串提取出来,得到“萨达姆”,排除“达姆”,最后把两个结果进行交集运算,得到“萨达姆”[9]。这两种方法,在时间和空间上开销较大,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的最长公共前缀(LCP,Longest Common Prefix),可以排除“萨达”。同时,利用上文的后缀信息来排除候选串,如果上文有相同的汉字,则不算作候选字串。如,“达姆”所在的后缀数组,其上文都为“萨”,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,利用后缀数组的 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》语料中提取出来的部分以“萨”和“达”开头的后缀数组,左图中可以直接提取出“萨达姆”、“萨达姆总统”等候选串,右图则显示出“达姆”的上文为“萨”或“阿”,可以直接排除掉“达姆”这个短串。而“阿达姆库斯”则可以通过“阿”开头的后缀数组提取出来。

上 文	后缀数组	LLCS	LCP
报道,伊拉克总统	萨达姆·侯赛因决定	0	3
任“维和”任务。	萨达姆下令释放约旦	0	3
,除非伊拉克总统	萨达姆停止干扰武器	0	3
5—7 级偏北风。	萨达姆呼吁联合国取	0	3
器核查小组人员。	萨达姆就海湾战争爆	0	3
记者招待会上感谢	萨达姆总统作出的上	0	5
旦囚犯。报道说,	萨达姆总统日前在巴	0	5
波苏瓦柳克当天向	萨达姆总统转交了叶	0	5
时作上述表示的。	萨达姆总统28日晚	0	5
激化矛盾克林顿称	萨达姆无权决定人选	0	5
立的。克林顿说:“	萨达姆无权决定谁来	0	5
和信仰进行自卫”。	萨达姆是在接见抗政的	0	3
)根据伊拉克总统	萨达姆的命令,首批	0	4
了上述决定。根据	萨达姆的这项决定,	0	4
在谈到伊拉克总统	萨达姆要求在6个月内	0	3
瓦柳克。俄特使向	萨达姆转交了叶利钦的	0	3
返回约旦。同时,	萨达姆还下令准许被	0	4
外国特委会合作。”	萨达姆还警告说,美国	0	4
国关系一度紧张。	萨达姆这项释放全部在	0	3
其立场更为明确。	萨达姆17日为纪念海	0	3
正龙)伊拉克总统	萨达姆17日呼吁联合	0	3
意愿留在伊拉克。	萨达姆17日在接见各	0	3
正龙)伊拉克总统	萨达姆29日表示,“	0	3
林顿:警告伊拉克	萨达姆:不希望战争	0	3

上 文	后缀数组	LLCS	LCP
,伊拉克总统萨	达姆·侯赛因决定	1	2
维和”任务。萨	达姆下令释放约旦	1	2
非伊拉克总统萨	达姆停止干扰武器	1	2
7 级偏北风。萨	达姆呼吁联合国取	1	2
查小组人员。萨	达姆就海湾战争爆	1	2
索瓦尔达斯·阿	达姆库斯以微弱优	1	4
国双重国籍。阿	达姆库斯当天表示	1	4
陶宛新总统。阿	达姆库斯现年71	1	4
招待会上感谢萨	达姆总统作出的上	1	4
犯。报道说,萨	达姆总统日前在巴	1	4
瓦柳克当天向萨	达姆总统转交了叶	1	4
上述表示的。萨	达姆总统28日晚	1	4
矛盾克林顿称萨	达姆无权决定人选	1	6
克林顿说:“萨	达姆无权决定谁来	1	6
和进行自卫”。萨	达姆是在接见抗政	1	2
据伊拉克总统萨	达姆的命令,首批	1	3
述决定。根据萨	达姆的这项决定,	1	3
到伊拉克总统萨	达姆要求在6个月	1	2
克。俄特使向萨	达姆转交了叶利钦	1	2
约旦。同时,萨	达姆还下令准许被	1	3
特委会合作。”萨	达姆还警告说,美	1	3
系一度紧张。萨	达姆这项释放全部	1	2
场更为明确。萨	达姆17日为纪念	1	0
)伊拉克总统萨	达姆17日呼吁联	1	0

图 2 1998 年 1 月语料中“萨达姆”的排序后缀数组示例

对于频次为 1 的字串,本文使用左右扩展法来获取。方法是利用后缀数组中频次为 1 的二字串进行左右扩展。每次扩展的二字频需为 1,否则后退一字。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。以“萨达”为出发点向左扩展,“诉萨”频次为 1,扩展为“诉萨达”;再往左扩展,由于“起诉”的频次大于 1,所以删除“起诉”,剩下“萨达”,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相应的元素“达姆”。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。即,多次词与单次词相连,且单次词内部的所有邻接二字对的频次也为 1。

后缀数组、最长公共前缀数组 LCP 和上文最长公共后缀 LLCS 的构造算法如下:

设 $S = C_1 C_2 \cdots C_i \cdots C_N$ 为一个由 N 个字符构成的字符串。$1 \le i \le N$, C_i 属于字符集 $\sum$。$A_i = C_i C_{i+1} \cdots C_N$ 为字符串 S 从字符 i 开始的后缀,可以得

到$\{A_1, A_2, \cdots, A_N\}$。按照字符集$\sum$的顺序进行排序,得到一个排序的二维数组 SA,即 $SA_1 < SA_2 < \cdots < SA_N$,“ < ”表示字符集中的先后顺序。数组 $LCP[1 \cdots N]$ 存放 SA_i 与 SA_{i-1} 或 SA_{i+1} 的公共最长前缀的长度(选择最大值),数组 $LLCS[1 \cdots N]$ 存放 SA_i 与 SA_{i-1} 或 SA_{i+1} 的上文公共最长后缀的长度(选择最大值)。

候选字串提取算法如下:

```
for(int i = 1; i < = N; i++){
//对于每一个后缀数组的元素
    if(LLCS[i] > 0)
    //如果上文相同的汉字个数 > 0,则跳过,不算候选串
        break;  //跳出本次循环
    else if(LCP[i] > 1){  //如果上文不同,且下文相同的汉字个数 > 1,则把该串加入 Hash 表
        HashTable- > insert(SA[i],LCP[i]);
        break;  //跳出本次循环
    }
    int m = 0,n = 0;  //m 为向左扩展的字数,n 为向右扩展的字数
    if(Search2gram(pp[i][0],pp[i][1]) = = 1){
    //如果上文不同,且 2 字频次为 1
        while((Search2gram(pp[i][-m], pp[i][-m - 1]) = = 1)
            &&(Search2gram(pp[i][-m - 1], pp[i][-m - 2]) = = 1))
        {  //依次向左扩展,条件是 Cm-1Cm 二字频次为 1,且 Cm-2Cm-1 二字频也为 1
            blind(pp[i][-m], pp[i][-m - 1]);  //屏蔽 Cm-1Cm 对应的 SA
            m++;  //左扩展字数加 1
        }
        while((Search2gram(pp[i][n], pp[i][n + 1]) = = 1)
            &&(Search2gram(pp[i][n + 1], pp[i][n + 2] = = 1))
        {//依次向右扩展,条件是 CnCn+1 二字频次为 1,且 Cn+1Cn+2 二字频也为 1
            blind(pp[i][n], pp[i][n + 1]);  //屏蔽 CnCn+1 对应的 SA
            n++;  //左扩展字数加 1
        }
        if(m + n > 0)  //如果左右扩展过了
            HashTable- > insert(SA[i],m + n + 1);
            //把该串加入 Hash 表
    }
}
```

3.3　互信息过滤

由后缀数组得到的字串需要经过一定的过滤,以提高精确率。文献[10]对 9 种常用的统计量进行了测试与分析,指出各统计量之间互补性不高。通常情况下,建议直接选用单个效果最好的互信息进行二字词的自动抽取。本文对于二元至四元字串采用条件熵推导出来的点互信息公式,对于五元以上的字串,由于公式过于繁琐,计算量过大,采用另一个简化公式。具体计算公式如下,

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

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

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

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

其中,$n \geq 1$,$f(C_1, \cdots, C_n)$ 是 n 元字串 $C_1, \cdots, C_n$ 在语料中出现的次数,N 是语料规模(总字数),$P(C_1, \cdots, C_n) = f(C_1, \cdots, C_n)/N$,$P(ac)$ 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,“3 000”、“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。

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“巴勒斯坦”,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。

图 3 人机交互式分词界面

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。

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。

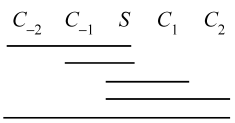

图 4 字串的外部查找模式

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 所在的前 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,其他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。

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。对于用户劳动量,以用户在进行交互时花费的总时间为依据,采用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。

表 1 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	149 886 字	新闻	北大	SIGHAN2005 ①
	简体	英雄出世	54 594 字	小说	北大	南师大 ②
	繁体	as_test	172 181 字	新闻	台湾	SIGHAN2005
近代汉语语料	繁体	红楼梦	718 388 字	小说	台湾	台湾“中研院” ③

① SIGHAN 2005 分词语料的下载网址:http://www.sighan.org/bakeoff2005/。

② 《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。

③ 《红楼梦》的分词标记文本惠蒙台湾“中研院”友情提供。该语料的特殊问题:在“中研院”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 ”。

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)		13 148	5 007	18 812	19 003
多字词(条)		11 672	3 738	17 246	15 921
交互情况	候选串(条)	5 926	3 376	9 089	15 713
	候选正确(条)	3 797	1 684	6 824	6 374
	正确率	0.640 7	0.498 8	0.750 8	0.405 7
	召回率	0.325 3	0.450 5	0.395 7	0.400 4
	F 值	0.431 5	0.473 4	0.518 2	0.403 0
	用户添加(条)	1 599	399	2 770	1 968
	得到总词表(条)	5 396	2 083	9 594	8 342
	总召回率	0.462 3	0.557 2	0.556 3	0.524 0
	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	0.857 6	0.846 9	0.801 9	0.935 4
	召回率	0.921 0	0.864 4	0.897 7	0.951 2
	F 值	0.888 2	0.855 6	0.847 1	0.943 2
分词 Baseline	未经人机交互	0.760 6	0.675 9	0.721 7	0.722 6
分词 TopLine	答案词表 + FMM	0.975 2	0.912 0	0.966 4	0.973 8

4.3　测试结果及分析

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 72 %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,得到的分词精度最高(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,在无人机交互条件下,系统的 F 值已达到 40 %以上;人机交互后的总召回率在 50 %左右,而且用户添加的词条仅占总词条的 1/3 以下,用户的劳动量并不是太大。不同的语料在正确率、召回率和用户添加词语的比例上有所差异,不过抽取的候选串的 F 值基本相近,系统得到的词(候选正确 + 用户添加)的总召回率是较为一致的,都保持在 50 %的水平上,其中的差异可能是由语料的特殊性造成的。然而,文本中依然有 50 %左右的多字词没有识别出来,其中绝大多数是出现 3 次以下的词语,说明系统在获取低频词语方面还需要改进,分词错误也主要是由未登录词引起的。

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。在实际使用时,系统还允许用户使用其他先验词表,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,系统可以把未切分串中的高频条目进行人工判定,从中提取出一些高频未登录词,还可以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所提升。由于使用该模块并不能直接得到未登录词,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 字符集为内核,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。在未登录词识别方面,重点解决了使用后缀数组时长短串覆盖问题和频次为 1 的字串的提取问题。

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。

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. 1996, 9(4):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,2002, 18(4):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[C]. Taejon, 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,2006, 20(1):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,2001, 15(6):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.

[10] 罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报, 2003, 17(3):9-14.