文章编号:1000-582X(2010)10-110-08

面向文本知识管理的自适应中文分词算法

冯　永,贺　迅,唐　黎,陈显勇,陈　贞
(重庆大学 计算机学院,重庆 400044)

摘　要:针对传统字典匹配分词法在识别新词和特殊词处理方面的不足,结合 2 元统计模型提出了面向文本知识管理的自适应中文分词算法——SACWSA。SACWSA 在预处理阶段结合应用有限状态机理论、基于连词的分隔方法和分治策略对输入文本进行子句划分,从而有效降低了分词算法的复杂度;在分词阶段应用 2 元统计模型,结合局部概率和全局概率,完成子句的切分,从而有效地提升了新词的识别率并消除了歧义;在后处理阶段,通过建立词性搭配规则来进一步消除 2 元分词结果的歧义。SACWSA 主要的特色在于利用"分而治之"的思想来处理长句和长词,用局部概率与全局概率相结合来识别生词和消歧。通过在不同领域语料库的实验表明,SACWSA 能准确、高效地自动适应不同行业领域的文本知识管理要求。

关键词:知识管理;文本处理;统计方法;自适应算法
中图法分类号:TP182　　**文献标志码**:A

Text knowledge management oriented adaptive Chinese word segmentation algorithms

FENG Yong, HE Xun, TANG Li, CHEN Xian-yong, CHEN Zhen
(College of Computer Science, Chongqing University, Chongqing 400044, P. R. China)

Abstract: To overcome the shortcomings of new word recognition and special word processing for the traditional dictionary-based matching algorithm in, text knowledge management oriented adaptive Chinese word segmentation algorithm (SACWSA) based on 2-gram statistical model is presented. . At the preprocessing stage, SACWSA applies finite state machine theory, conjunction-based partition method and divide conquer strategy to partition long sentences in input text into sub-sentences, which reduces the algorithm complexity effectively. At the word segmentation stage, 2-gram statistical model is employed and combined with partial probability and overall probability to partition the sub-sentences into words, which improved the recognition rate of new words and eliminated ambiguity. At the post-processing stage, the matching rules of part-of-speech are established to eliminate ambiguity of 2-gram word segmentation results further. The innovations of SACWSA include dealing with the long sentences and long terms with the idea of 'Divide and Conquer'; while combining the partial probability and overall probability to identify new words and eliminate ambiguity. Experimental results on text corpus of different fields show that SACWSA can adapt to different text knowledge management requirements in different fields accurately, efficiently and automatically.

Key words: knowtl edeg management; text processing; statistical methods; adaptive algorithms

收稿日期:2009-05-10
基金项目:重庆市自然科学基金资助项目(2008BB2183);中央高校基本科研资助项目(DJIR10180006);"211 工程"三期建设资助项目(S-10218);中国博士后科学基金资助项目(20080440699);国家科技支撑计划资助项目(2008BAH37B04);国家社会科学基金"十一五"规划教育学重点课题(ACA07004-08)。
作者简介:冯　永(1977-),男,重庆大学副教授,主要从事知识发现、数据挖掘、自然语言处理等方向研究,(Tel)13983980003;(E-mail)fengyong@cquedu.cn。

中文分词是机器翻译、分类、主题词提取以及信息检索的重要基础[1]。面向文本知识管理的中文分词有着强烈的目的性，主要考察其是否有助于提高知识文本信息检索的准确度，以及对行业知识新词的识别（包括人名、地名、组织名和其他不在词典中的术语、俚语或网络用语的识别）能力和歧义的解决（包括交叉歧义和组合歧义的解决）能力[2-3]。传统的字典匹配分词方法虽然有着较高的准确性，但是其分词性能受限于词典的完备性，从而无法适应现实日益发展的领域知识管理需求[4]。从统计理论出发，提出了一种能自适应中文语料和领域的分词方法，具有较强的适应性和实用性。

1　中文分词技术

目前国内外比较权威的汉语分词系统所采用的分词方法，主要有 3 种类型[5]：基于字典匹配的分词法、基于语料统计的分词法、规则和统计结合的方法。

1.1　基于词典匹配的分词法

从现有文献分析[6-7]，取得较好效果的方法主要有最大匹配法、逆向最大匹配法、双向匹配法、最佳匹配法。

最大匹配的原理简单，易于在计算机上实现，时间复杂度也比较低。但是，最大词长难于确定；逆向切词的正确率比最大匹配要高一些，但是需要配置逆序的切词词典，与自然语言习惯不一致，维护不便；双向匹配对于纠错有一定的效果，但是对于正、逆向扫描的结果虽一致但实际上讲切分不正确的字段仍没有强有力的处理手段；最佳匹配通过缩短查询词典的时间，加快分词的速度，从而降低了分词的时间复杂度。

基于词典匹配的分词法，实现简单，实用性强，但该分词法的最大的缺点就是词典的完备性不能得到保证。

1.2　基于语料统计的分词法

基于统计的方法的基础是利用汉字同时出现来组成有意义词的概率。设 W 为待分词的句子，S 为分词后的结果，统计的方法就是把找“正确的”解转化为找最可能的解，即找最大似然估计 S，使得在 W 下的后验条件概率最大。根据 Bayes 公式，可以将后验条件概率转化为先验概率进行计算。于是，统计方法就转化为估计概率和构造算法以求对应的解。作为一种十分简单而且非常有效的语言模型，通过一个足够大的语料库进行有监督或者无监督的学习，可以用一阶马尔科夫假设和独立性假设来进行分词处理[8-9]。

基于语料统计的分词方法有许多优点[10-11]：未登录词的影响降低了，只要有足够的训练文本就易于创建和使用。

1.3　规则和统计结合的方法

部分分词算法采用规则和统计相结合的办法，可以降低统计对语料库的依赖性，充分利用已有的词法信息，同时弥补规则方法的不足[12-16]。

2　自适应中文分词算法的理论基础

2.1　文本知识管理对中文分词的要求

1）准确性。准确性是分词算法设计的核心指标，它是“正确切分的词数/切分出来的所有词数”。

2）高效性。在所有文本处理的相关软件系统中，分词是共同、频繁而基础性的操作。这部分操作对于应用系统来讲，时间和系统开销越少越好，特别在大文本加载时尤为重要。

3）适用性。分词产生的结果是为某个具体应用服务的，对不同领域的文本，分词系统往往获得的准确率不尽相同。好的分词系统应能在处理不同领域的文本知识时都能够达到可以接受的准确率和性能，这是论文设计算法的主要目标。

2.2　N 元（N-gram）统计模型原理

N 元语法的基本思想是：1 个单词的出现与其上下文中出现的其他单词密切相关。1 个句子可以看成 1 个有联系的字符串序列，可以是字序列，也可以是已知的词构成的词序列。对于 1 个句子 $w_1w_2\cdots w_k$ 的出现概率用 $P(W)$ 来表示，有

$$P(W)=P(w_1w_2\ldots w_k)=P(w_1)P(w_2|w_1)$$
$$P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})=\prod_{i=1}^{n}P(w_i|w_1w_2\ldots w_{i-1})。\quad(1)$$

从字的角度来看，该模型认为第 k 个词的出现与前面 $k-1$ 个词相关。为了预测 w_k 的出现概率，就必须知道前面所有词的出现概率，其计算过于复杂甚至是不可能的。

由此可见，N-gram 方法实际把分词问题转化为求最佳的分词组合 $w_1w_2\cdots w_k$，使得 $P(W)$ 的值最大。

如果假设 w_k 只与其前面出现的 $n-1$ 个词有关，就是 N 元模型。比如只与前面的两个词有关，则称该语言模型是三元模型。公式简化为

$$P(W) \approx P(w_1)P(w_2 \mid w_1)\prod_{i=3}^{k} P(w_i \mid w_{i-2}w_{i-1})。 \quad (2)$$

公式中的概率参数均可以通过大规模语料库来进行计算

$$P(w_i \mid w_{i-2}w_{i-1}) \approx \frac{\text{count}(w_{i-2}w_{i-1}w_i)}{\text{count}(w_{i-2}w_{i-1})}。 \quad (3)$$

其中，count(L)表示字串 L 在整个语料库中出现的累计次数。

2 元语法模型，也叫一阶马尔科夫链。即

$$P(W) \approx \prod_{i=1}^{k} P(w_i \mid w_{i-1})。 \quad (4)$$

研究使用的 2-gram2 元模型。

2.3 N-gram 方法对分词性能的影响

在 N-gram 方法中，为了预测 w_k 的出现概率，就必须知道前面所有词的出现概率，因此，影响其分词性能的原因有以下几个方面

1)语料库的代表性和完备性。字串 $L(w_k, w_1w_2\cdots w_{k-1})$在语料库中出现的次数，直接影响到计算 $P(w_k|w_1w_2\cdots w_{k-1})$，因此，训练语料库中的这些字串是否有代表性和完备性，将直接影响到它对新词的识别能力。

2)最大词长和最大句长。最大词长是 N-gram 方法所选择的 N，最大句长是待分词的句子的最大长度。由于 N-gram 方法需要组合句子的各种分解方式，比较各个方式的 $P(W)$值，句长对其影响尤为显著，因此其时间复杂度随句长的增加呈指数增加。研究表明[12]，当句长＞30，词长为 4 时，使用 N-gram 方法已经无法有效求解。

3)算法搜索方式。在 N-gram 方法组合的句子的各种分解方式中找到最优解，将这些可能的分解方式组成一颗树，而对这颗树就可以使用深度优先、宽度优先、长词优先等搜索方式。

3 自适应中文分词算法 SACWSA

基于上述要求和理论基础，论文提出了一种能够很好适应多种语料信息的分词算法 SACWSA (self-adaptive chinese word segmentation algorithm)，并且时间和精度都能够满足实际文本知识管理系统的应用需要，能较好地解决新词和特殊词的识别问题。

3.1 算法的基本流程

算法中首先根据预处理规则对输入文本进行子句划分，接着利用 2-gram 方法对各个子句进行分词，然后进行子句合并处理，最后根据后处理规则得到分词结果。

3.2 算法的预处理方法

预处理的主要目的，是将长句进行划分为多个子句，从而降低分词算法的复杂度。论文的方法有 3 种。

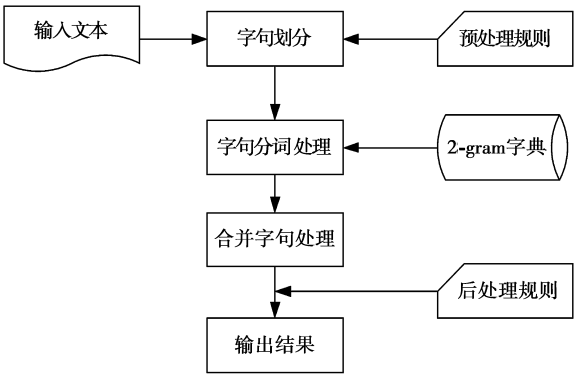

图 1 算法基本流程图

1)利用有限状态机识别待分词文本中最常见的数字、日期以及域名等并以它们为标志将句子划分为子句。例如，利用有限状态机识别年份上图给出了 1 个较为复杂的例子，在针对具体语料进行处理时，可以进行简化。针对数字识别的 FSA 更为简单，论文在此不赘述。论文对 Sogou 实验室(SOHU 公司)提供的语料数据进行了预处理实验对比，利用预处理有效缩短了长句(论文指大于 30 字，下同)的概率，从 3%降低到 1.4%。实验中利用这种方法进行长句划分的例子如下

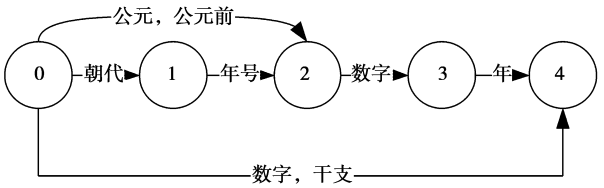

图 2 识别年份的有限状态机

“沙特阿拉伯石油和矿产资源大臣纳伊米 9 日预测说石油价格到 2010 年将会稳定”被预处理为：“沙特阿拉伯石油和矿产资源大臣纳伊米/9 日/预测说石油价格到/2010 年/将会稳定”。

2)对于用上述方法无法处理的长句，论文定义一个连词集合，以基于连词分隔的方式进行处理，这样可以进一步将长句的概率下降到 2‰。只使用常出现在语句中部的连接词作为有意义的语句划分

词。例如,连词集合{暨,甚至,就是,和,然而,就,且},划分的长句例如

“像最近上市的氨酚曲马多片/就是/由阿片类/和/非甾体类使用最久的盐酸曲马多/和/对乙酰氨基酚组合在一起的复方产品”。

3)对于以上 2 种方法都无法处理的长句,则采用分治的方法,直接将句子划分为 t 个子句(假设除最后一个子句外的前面子句长度为 k),先对各个子句进行分词,最后归并其结果。其思路如图 3 所示

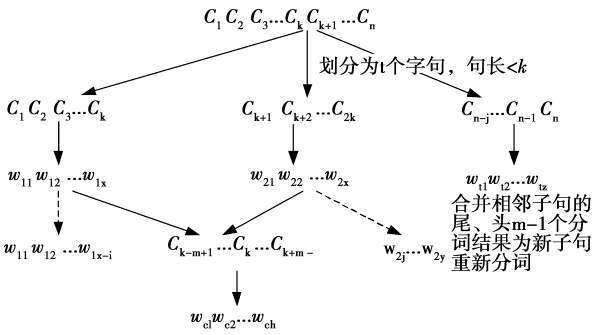

图 3　“分而治之”长句划分思想

在实际应用中,例如

长句“我真的希望看到越来越多的中国人能够出现在世界最高水平的联盟中证明我们中国人的实力”被直接划分为“我真的希望看到越来越多的中国人能够出现/在世界最高水平的联盟中证明我们中国人的实力”2 个子句。

3.3　子句 2-gram 分词算法

使用预处理方法将待处理文本中的长句划分为子句后,对每个子句采用 2-gram 算法,步骤如下

输入:

经过预处理后的文本文件 $s=s_1s_2\cdots s_n$;$s_i=c_{i1}c_{i2}\cdots c_{ij}$ 其中 c_{ij} 均为单字;

从语料库加工的词频字典。

处理流程:

Step1:用二级 Hash 表加载词频字典并做数据平滑。

Step2:使用词长优先获得二元切分路径。

Step3:使用深度优先算法选择最优路径。

输出:

由 s_i 最优路径分词结果 $w_{i1}w_{i2}\cdots w_{ik}$ 组成的 s 分词结果。

算法的部分关键代码(伪代码实现)如图 4 和 5。

```
public static void add(String prewd,String currwd){
        String key=prewd;
        String curr=currwd;
        boolean bb=HMap.containsKey(key);
         if(bb==false){
            HashMap hm=new HashMap();
            hm.put(key,new Integer(1));
                        hm.put(purr,new Integer(1));
           HMap.put(key,hm);
         }
         else
         {
            HashMap temp=(HashMap)HMap.get(key);
            int count=((Integer)temp.get(key)).intValue(+1);
            temp.put(key,new Integer(count));
             if(temp.containsKey(curr))
             {
                int value=((Integer)temp.get(curr)).intValue()+1;
                            temp.put(curr,new Integer(value));
             }
             else
                temp.put(curr,new Integer(1));
            HMap.put(key,temp);
         }
    }
```

图 4　代码片段 A:从训练语料中建立 2-gram 模型(加载词频字典)

```
public static void add(String prewd,String currwd){
        String key=prewd;
        String curr=currwd;
        boolean bb=HMap.containsKey(key);
         if(bb==false){
            HashMap hm=new HashMap();
            hm.put(key,new Integer(1));
                        hm.put(purr,new Integer(1));
           HMap.put(key,hm);
         }
         else
         {
            HashMap temp=(HashMap)HMap.get(key);
            int count=((Integer)temp.get(key)).intValue(+1);
            temp.put(key,new Integer(count));
             if(temp.containsKey(curr))
             {
                int value=((Integer)temp.get(curr)).intValue()+1;
                            temp.put(curr,new Integer(value));
             }
             else
                temp.put(curr,new Integer(1));
            HMap.put(key,temp);
         }
    }
```

图 5　伪代码片段 B:全切分最佳路径选择

3.4　算法的关键问题分析

1)词频字典的准备。利用《人民日报》1998 年 1 月份标注的语料来设计 2-gram 词典库,该语料库由于同时有词性标注,论文算法后处理的时候使用词

性搭配进行交差歧义的判别。形成以词为索引的词频表,其出现的频率为值。同时根据该语料库构建两个词之间的搭配频度。语料库中的词汇量为55 000个,其中名词12 000余个,出现了460 000种双词搭配使用的情况。

目前语料库中词性的标记集里除了有26个基本词类标记,即名词n、时间词t、处所词s、方位词f、数词m、量词q、区别词b、代词r、动词v、形容词a、状态词z、副词d、介词p、连词c、助词u、语气词y、叹词e、拟声词o、成语i、习惯用语l、简称j、前接成分h、后接成分k、语素g、非语素字x、标点符号w。此外从语料库应用的角度,还增加了专有名词(人名nr、地名ns、机构名称nt、其他专有名词nz)等其他标记,总共使用了39个标记。

从语料库中很容易构建词性搭配的2元词典。

2)由于数据稀疏的存在,为了避免在计算后续每种切分的$P(W)$时出现0概率,应对数据进行平滑。由于算法过程比较的是$P(W)$的大小,平滑方法对比较结果影响很小,采用Add-delta平滑技术[17],对不同平滑技术未做比较研究。

3)对输入的$s_i=c_{i1}c_{i2}\cdots c_{ij}$,简记为$c_1c_2\cdots c_n$,使用2元迭代法选取切分路径。即首先以2元切分作筛选,筛选后,合并选取的2元词,把它们作为字进行下一轮迭代,3次迭代就可以发现8个字的词,收敛速度较快。第一次迭代时,候选词为$\{c_1c_2,c_2c_3,c_3c_4,\cdots,c_{n-1}c_n\}$,可以直接从语料库字典中计算比较$P(c_1c_2)$和$P(c_2c_3)$决定选择$c_1c_2$还是$c_2c_3$。第一次迭代可以把所有的单字词和双字词识别出来,然后再将这个结果作为输入,进行2元筛选。

4)选择最佳切分时,统计值只用于比较,在保持统计值计算相同的前提下,论文采用先比较局部概率,当他们相等时再比较全局概率。局部概率是它在输入的文本文件s中出现的概率,全局概率是在整个语料库中出现的概率。语义的局部特征决定了同一文本中的词会集中表达相同或者相似的语义。在训练集中没有出现的词而在特定文本中反复出现,反而更有代表性。对局部概率的计算仍然统计频率,需要设定阈值来决定是否使用,笔者把这个阈值设定为"词频+转移频率",这里的转移频率是指词与其后续不同词出现的频率。论文将这个阈值设定为2。

例如,对于最常见的生词识别,假设输入的句子是"奥巴马登上了飞机。奥巴马刚与希拉里会面研究了竞选搭档的问题"。第一次迭代时,由于"奥"和"巴马"的转移频率(1)<"巴马"的转移频率(2),所以,"奥巴马"被切分为"奥/巴马";第二次迭代的时候,因"奥巴马"出现了两次并且转移频率也为2,所以"奥巴马"被组合成为一个词。

当前后2个2元词都存在时,也就是存在歧义的时候,需要比较3个2元词的局部概率和全局概率,以确定中间的这个二元词是否被切分。比如,"郑重要求",会被切分为"郑重/要求"而不是"郑/重要/求"。除此之外,论文在处理规则中还采用了词性搭配方式检查交差歧义。

由于局部概率的采用,使得论文提出的算法对待处理文件具有很强的适应性。虽然对于仅出现一次的新词或歧义,论文的算法不能全部识别,但是由于反映主题的新名词往往会在上下文中反复出现,因此,实际应用时优势十分明显。特别在不同的行业领域应用时,对于领域知识的专有名词的识别率十分明显。

3.5 算法的后处理规则

论文的后处理主要是用简单的词性搭配规则对2-gram切分结果进行歧义发现和处理。由于2-gram分词结果的词性重新组合为1个2元模型,设定词性搭配阈值进行筛选,发现可能产生歧义的近邻2元词,并重新进行切分。

例如:

输入:"太平淡的故事",由于在语料库中"太平"的词频比"平淡"的词频要高,2-gram分词结果为"太平/淡/的/故事"(a+a,形容词+形容词)。词性搭配检查发现,形容词+形容词的搭配方式小于给定的阈值,需要对其重新进行切分,从而得到正确的结果:"太/平淡/的/故事"。

4 实验与分析

实验是在CPU为Intel Core2 Duo 1.80 GHZ的机器上完成的,内存为1 G,操作系统为Windows Server 2003。测试精度和速度时所使用的待处理文本数据为SOGOU实验室文本分类语料库。该语料库包含9类文件,财经,IT,健康,体育,旅游,教育,招聘,文化,军事,每一类文件1 990个。同时还使用了《人民日报》1998年1月语料库进行封闭测试。

为了测试论文提出算法的适应性,还选择了3个工程项目(3个不同领域,其主题性非常强)的待处理文本各20篇,并对分词结果进行了人工评测。

4.1 测试的分词结果举例

1)使用SOGOU语料库进行分词的结果摘录

原文:

近日,美国医药保健管理协会(PCMA)的一份

评估报告指出,专利药厂家就老年人经常使用的专利药出台了一系列措施,试图限制仿制药厂商产销该类产品,通过诉讼、游说立法和专利成分陷阱等手段,力求最大程度延缓仿制药上市或进入联邦医疗保险体系。

分词结果:

近日/美国/医药/保健/管理/协会/(PCMA)/的/一/份/评估/报告/指出/专利/药厂/家/就/老年人/经常/使用/的/专利/药/出台/了/一/系列/措施/试图/限制/仿/制药/厂商/产销/该类/产品/通过/诉讼/游说/立法/和/专利/成分/陷/阱/等/手段/力求/最/大/程度/延缓/仿制/药/上市/或/进入/联邦/医疗/保险/体系。

2)使用人民日报语料库进行分词的封闭测试摘录

原始语料为:

农业生产再次获得好的收成,企业改革继续深化,人民生活进一步改善

对外经济技术合作与交流不断扩大

民主法制建设、精神文明建设和其他各项事业都有新的进展

我们十分关注最近一个时期一些国家和地区发生的金融风波,我们相信通过这些国家和地区的努力以及有关的国际合作,情况会逐步得到缓解

分词后的结果为:

农业/生产/再次/获得/好/的/收成/企业/改革/继续/深化/人民/生活/进一步/改善

对外/经济/技术/合作/与/交流/不断/扩大

民主/法制/建设/精神文明/建设/和/其他/各项/事业/都/有/新/的/进展/

我们/十分/关注/最近/一个/时期/一些/国家/和/地区/发生/的/金融/风波/我们/相信/通过/这些/国家/和/地区/的/努力/以及/有关/的/国际/合作/情况/会/逐步/得到/缓解

语料库中的分词结果(标准答案):

农业/生产/再次/获得/好/的/收成/企业/改革/继续/深化/人民/生活/进一步/改善

对外/经济/技术/合作/与/交流/不断/扩大/

民主/法制/建设/精神文明/建设/和/其他/各项/事业/都/有/新/的/进展

我们/十分/关注/最近/一个/时期/一些/国家/和/地区/发生/的/金融/风波/我们/相信/通过/这些/国家/和/地区/的/努力/以及/有关/的/国际/合作/情况/会/逐步/得到/缓解。

4.2　算法的分词正确率

由于分词结果的正确率需要人工评测,其工作量巨大且繁琐,与人工评测的标准和测试文本集有关,所以商用的各个分词系统的性能之间无法进行准确对比。目前,ICTCLAS 是能够免费获取使用的最好分词系统,以其为对比的参照系统,在封闭测试中,人工选择了 20 个文本进行评测计算。Sogou 语料库的开放测试从各类选出 2 个文本共 18 个文本进行人工评测。

参照通用的标准,评测指标定义如下

$$\text{切分准确率}(P)=\frac{\text{正确切分的词语数}}{\text{切分出来的所有词语数}};$$

$$\text{切分召回率}(R)=\frac{\text{正确切分的词语数}}{\text{答案中的所有词语数}}。$$

实验是在最大句长为 28,最大词长为 8 的情况下,得到下面的切分准确率和切分召回率。

表 1　语料库测试结果人工评测对比表

	P(Sogou 语料库)	R(Sogou 语料库)	P(封闭)	R(封闭)
SACWSA	0.96	0.956	0.97	0.968
ICTCLAS	0.96	0.95	0.96	0.952

提出的 SACWSA 算法在封闭测试和用 Sogou 语料库测试与 ICTCLAS 测试结果的比较,均能达到很好的商用性能。有文献报道 ICTCLAS 在部分领域能够达到 97%以上的切分准确率,与论文的人工评测结果基本接近。

4.3　算法的适应性测试

论文选择了 3 个领域性强的知识文本来进行适应性测试,各 20 个长度为 3 000～5 000 字的文本文件。领域性强主要体现在生词较多。例如

1)农业领域“无土栽培”文本知识摘录。

“理论上,无土栽培可以种植各种可在土壤中生长的作物,包括……。花卉类包括……、唐菖蒲、曼丽榕、九里香、……等。”

在这段文章中,出现了多个在人民日报语料库中没有出现的新词,比如“唐菖蒲”、“曼丽榕”、“九里香”。而是否能将这些新词准确切分出来,将直接影响到知识管理后续的全文检索、分类等重要应用。

2)税收领域“企业所得税”知识文本摘录.

“《中华人民共和国企业所得税法》(以下简称“新税法”)已于 2008 年 1 月 1 日起施行,为保证企业所得税……”

在这段文章中,“企业所得税”是 1 个组合歧义的代表,在人民日报语料库中,这个词被分为“企业”

和“所得税”2 个词。是否被切分为 1 个词对于分类精度、检索精度都会带来显著的影响。

3)钢铁行业“信息情报”文本知识摘录。

“一直被市场视作钢材价格“风向标”的宝钢价格……,普冷下调 300 元/t,电镀锌下调 100 元/t……”

在这段文章中,也频繁出现人民日报语料库中没有的专业词汇,比如“普冷”、“电镀锌”等。

实验中,引入了未登录词识别准确率和未登录词识别召回率两个评测指标。其定义如下:

$$\text{未登录词识别准确率}(P_{oov})=\frac{\text{识别正确的未登录词数}}{\text{识别出来的未登录词数}};$$

$$\text{未登录词识别召回率}(R_{oov})=\frac{\text{识别正确的未登录词数}}{\text{答案中的未登录词数}}。$$

其中的答案是人工评测,分别由领域知识专家参与制定。实验结果如表 2

表 2 算法适应性测试结果

	P_{oov} (SACWSA)	R_{oov} (SACWSA)	P_{oov} (ICTCLAS)	R_{oov} (ICTCLAS)
农业领域	0.98	0.96	0.44	0.41
税务领域	0.98	0.95	0.54	0.50
钢铁情报	0.96	0.97	0.48	0.36

实验结果表明:ICTCLAS 在领域性较强的文本分词中对未登录词的识别率较差,这与其他研究所表明的情况一致。而提出的 SACWSA 算法,在 3 个不同领域的新词识别过程中,显示出较强的适应能力。其主要原因在于局部概率的引入,对于那些出现 2 次以上的新词能够很好地加以识别。

4.4 不同最大句长和词长对分词算法的影响

分词的效率主要体现在系统的运行速度上,运行速度对于大规模工程应用,特别是对大文本加载、Internet 情报检索等应用特别需要。由于各个系统对算法的实现采用了不同的工具和平台,比如有的使用 C 语言实现,也有使用 Java、VB 等语言实现,运算速度不尽相同。同时,运算速度也与具体的运行环境有关。

有文献报道 ICTCLAS 的分词速度在 800 k/s～2 M/s,其他若干算法和系统的分词速度也在 500 k/s～1 M/s。考虑到工程应用的实际情况对可移植性要求也较高,用 JAVA 实现,切分速度在不同的词长和句长下约为 300～700 k/s。实现的系统在具体的多个工程应用中,一般分词算法的速度只要不低于 300 k/s 就能够满足需求。

1)不同最大句长和最大词长对算法效率的影响

从“N-gram”算法本身分析,其原理在于组合句子的各种分解方式,比较各个方式的 $P(W)$值,句长对其影响尤为显著,其时间复杂度随句长的增加呈指数增加。由于提出的 SACWSA 算法的预处理方法对长句进行了拆分,句长主要影响到预处理的效率,而最大词长关系到切分路径的选择,论文使用 2 元迭代,收敛速度较快,不会造成太大的效率差别。用 SOGOU 语料库进行了实验测试,在最大句长和最大词长不同的搭配下得到分词的效率。图 6 是不同的最大句长和词长下的分词效率实验结果。

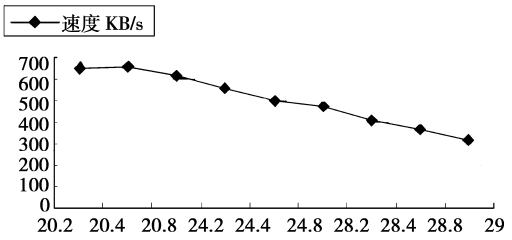

图 6 不同最大词长和最大字长下的分词效率

实验结果表明:句长是制约 SACWSA 算法的主要因素。其原因在于,虽然论文在预处理中将长句按规则切分成了多个子句,但是多个子句在最后合并的时候要重新进行组合分词,因此影响到了执行效率,划分的子句越多,这部分的处理时间消耗就比较多。实验结果也表明,由于切分路径的迭代速度很快,相同句长下不同词长对算法效率的影响相对较小。

2)不同最大句长和最大词长对准确率的影响

表 3 最大句长和最大词长对准确率的影响

最大句长	最大词长(迭代次数)	准确率/%
20	2	75
20	4	85
20	8	95.5
24	2	77
24	4	86
24	8	94.5
28	2	76.5
28	4	90
28	8	95

在实验中,对最大词长的控制,实际上反应了算法对切分路径选择的迭代次数。由于采取了对长句的预处理,所以句长对算法准确率的影响不大,实验结果也表明了这一点。而迭代次数对正确率的影响

显著,因为这制约了对长词的发现能力,虽然实验结果在很大程度上受到语料和人工评测答案的影响而不一定具备代表性,但实验结果仍表明,3 次迭代就可以达到较高的准确率。

5 结　语

论文结合 2 元统计模型的优势提出了面向文本知识管理的自适应中文分词算法—SACWSA。该算法利用了"分而治之"的思想对制约分词效率的长句问题进行了处理,利用局部概率和全局概率相结合来识别生词,利用词性搭配进行歧义消除。所提算法克服了传统的字典匹配分词法和一般统计分词法的弱点,使分词处理歧义和识别新词方面有较大的改善。通过对 3 个主题性强的领域文本进行的分词实验,说明算法对领域和语料信息的适应性较强,可以较广泛地应用于各个工程应用场景。

参考文献:

[1] GAO J F, WU A D, LI M. Adaptive Chinese word segmentation [C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. [s. l.]:ACL2004, 2004: 462-469.

[2] ZHANG M Y, LU Z D, ZOU C Y. A Chinese word segmentation based on language situation in processing ambiguous words[J]. Information Sciences, 2004, 162(3-4):275-285.

[3] WANG X J, QIN Y, LIU W. A search-based Chinese word segmentation method[C]. Proceedings of the 16th International World Wide Web Conference, 2007:1129-1130.

[4] WANG H S, CUI M M. A Chinese word segmentation based on machine learning [C]// Proceedings of the 1st International Workshop on Education Technology and Computer Science. [S. L.] ETCS 2009, 2009, 2:610-613.

[5] HONG C M, CHEN C M, CHIU C Y. Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems [J]. Expert Systems with Applications, 2009, 36(2): 3641-3651.

[6] ZENG D, WEI D H, CHAU M, et al. Chinese word segmentation for terrorism-related contents[J]. Lecture Notes in Computer Science, 2008, 5075:1-13.

[7] LUO X G, LUO J, XIE Z. The research of chinese automatic word segmentation in hierarchical model dictionary binary tree [C]//Proceedings of 1st International Workshop on Database Technology and Applications. [s. l.]: DBTA 2009, 2009: 321-324.

[8] 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.
FENG CHONG, CHEN ZHAO-XIONG, HUANG HE-YAN, et al. Active learning in Chinese word segmentation based on multigram language model[J]. Journal of Chinese Information Processing, 2006,20(1):50-58.

[9] 曹勇刚,曹羽中,金茂忠,等. 面向信息检索的自适应中文分词系统[J]. 软件学报,2006,17(3):356-363.
CAO YONG-GANG, G CAOYU-ZHON, JIN MAO-ZHONG, et al. Information retrieval oriented adaptive Chinese word segmentation system [J]. Journal of Software, 2006,17(3):356-363.

[10] YANG, C C, LI K W. A heuristic method based on a statistical approach for Chinese text segmentation[J]. Journal of the American Society for Information Science and Technology,2005, 56(13): 1438-1447.

[11] FU G H, KIT C Y, WEBSTER J J. Chinese word segmentation as morpheme-based lexical chunking[J]. Information Sciences, 2008, 178(9): 2282-2296.

[12] 张华平,刘群. 基于 N-最短路径方法的中文词语粗分模型[J]. 中文信息学报,2002,16(5):1-7.
ZHANG HUA-PING, LIU QUN. Model of chinese words rough segmentation based on N-Shortest-Paths method[J]. Journal of Chinese Information Processing, 2002,16(5):1-7.

[13] GOH C L, MASAYUKI A, MATSUMOTO Y I. Chinese word segmentation by classification of characters[J]. Computational Linguistics and Chinese Language Processing,2005,10(3):381-396.

[14] WANG Z R, LIU T. Chinese unknown word identification based on local bigram model [J]. International Journal of Computer Processing of Oriental Languages, 2005, 18(3): 185-196.

[15] XIONG Y, ZHU J. Feature study for improving Chinese overlapping ambiguity resolution based on SVM[J]. Journal of Southeast University (English Edition), 2007, 23(2):179-184.

[16] 宋彦, 蔡东风, 张桂平, 等. 一种基于字词联合解码的中文分词方法[J]. 软件学报, 2009, 20(9): 2366-2375.
SONG YAN, CAI DONG-FENG, ZHANG GUI-PING, et al. Approach to Chinese word segmentation based on character-word joint decoding [J]. Journal of Software, 2009, 20(9): 2366-2375.

[17] CHEN S F, GOODMAN J T. An empirical study of smoothing techniques for language modeling [R]. Harvard University, Tech Rep :TR210298, 1998.

(编辑　侯　湘)