◎信号处理◎

# 基于排序集成的哈萨克语固定短语抽取

桑海岩[1,2],古丽拉·阿东别克[1,2],孙瑞娜[3],陈 莉[1,2]
SANG Haiyan[1,2], Gulia·ALTENBEK[1,2], SUN Ruina[3], CHEN Li[1,2]

1.新疆大学 信息科学与工程学院,乌鲁木齐 830046
2.国家语言资源监测与研究中心少数民族语言中心 哈萨克和柯尔克孜语文基地,乌鲁木齐 830046
3.新疆财经大学 统计信息学院,乌鲁木齐 830046
1.College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China
2.The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center Minority Languages, Urumqi 830046, China
3.College of Statistical Information, Xinjiang University of Finance and Economics, Urumqi 830046, China

**SANG Haiyan, Gulia·ALTENBEK, SUN Ruina, et al. Rank aggregation-based Kazakh fixed phrases extraction. Computer Engineering and Applications, 2014, 50(21):205-209.**

**Abstract:** Phrase extraction plays a key role in text information understanding, such as automatic text classification, topic extraction, and analysis of patent search, etc. As the part of phrase research, the fixed phrase extraction has important practical significance on natural language processing tasks including the lexicographer. The Kazakh is agglutinative language, rich in inflections. These characteristics of the Kazakh bring certain difficulties to fixed phrase extraction. This paper proposes a general fixed phrase extraction algorithm. The algorithm considers the fixed phrase extraction as a scheduling problem, uses C-value, mutual information and log-likelihood statistics to extract and schedule, and presents a new rank aggregation method to obtain a scheduling result set. The experimental results indicate that the algorithm gets higher accuracy compared with popular signal extraction algorithms.

**Key words:** natural language processing; fixed phrases; rank aggregation; mutual information; log-likelihood; C-value

**摘 要:**短语抽取是文本自动分类、主题提取及专利检索分析等文本信息理解等工作中都要应用到的一项关键技术。固定短语抽取作为短语研究的一部分,对短语标注、辞典编撰等自然语言处理任务都具有重要的现实意义。哈萨克语是黏着语,词形变化丰富,这些特点给哈语固定短语的抽取带来了一定的困难。提出一个总体的固定短语抽取算法,把固定短语抽取看作一个排序问题,使用C-value、互信息和log-likelihood进行抽取排序,并设计了一个新的排序集成方法对抽取的结果进行集成。实验分析结果表明,与单独的抽取算法比较,该算法达到了更高的准确率。

**关键词:**自然语言处理;固定短语;排序集成;互信息;似然比;C-value算法
**文献标志码:**A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1211-0373

## 1 引言

短语抽取[1]是在文本自动分类、主题提取及专利检索分析等文本信息理解等工作中都要应用到的一项关键技术。固定短语抽取作为短语研究的一部分,对短语标注、辞典编撰等自然语言处理任务都具有重要的意义。

**基金项目:**国家自然科学基金(No.61063025);新疆多语种信息技术重点实验室开放项目(No.049807)。

**作者简介:**桑海岩(1982—),男,硕士,CCF会员,主要研究领域为自然语言信息处理;古丽拉·阿东别克(1962—),女,教授,博士生导师,主要研究领域为自然语言信息处理,人工智能等;孙瑞娜(1982—),女,讲师,主要研究领域为人工智能;陈莉(1988—),女,硕士,主要研究领域为自然语言处理。E-mail:sang_haiyan@163.com

**收稿日期:**2012-11-30 **修回日期:**2013-03-25 **文章编号:**1002-8331(2014)21-0205-05
CNKI**网络优先出版:**2013-05-03,http://www.cnki.net/kcms/detail/11.2127.TP.20130503.1708.011.html

哈语短语同汉语短语有相近概念,两个或两个以上的实词按照一定的结构规则组合而成的语言单位叫短语[2]。哈萨克语属于阿尔泰语系突厥语族的克普恰克语支,拼音文字,是黏着语言类型,有着高度丰富的形态变化。组成短语的词不仅要受到结构规则的制约而且又受语法关系的制约,主要表现在不同的语境下短语中词的词缀形态的改变。此外哈语中还含有丰富的曲折短语。曲折短语是指含有发生内部曲折词的短语,而词的内部曲折是指因为语法或发音的需要而发生的语音交替现象,这与汉语短语有很大区别。上述这些特点对哈语短语抽取带来了一定困难。哈语短语从稳定性上讲可以分为固定短语和自由短语[3]。固定短语是历史上固定下来的,在句子中作为一个单词使用,多为成语、熟语等。自由短语是由语义上能够搭配的两个或两个以上实词带入某种结构关系的词组模式得出的语言片段,词之间的组合比较自由,包括名词性短语,动词性短语等。本文中所说的固定短语是指经常在一起使用的表达一个完整意义的实词组合,包括了大量的成语、熟语以及实体名和专业术语等。

## 2 研究现状

短语抽取主要有两大方法:一是知识工程方法;二是统计方法[4]。知识工程方法要求编制规则的知识工程师对领域知识有深入的了解,而基于统计的方法则不需要。基于统计的方法中,目前最具有代表性的是log-likelihood[5]方法、互信息方法[6]、C值[1,7]和*N*-gram方法,前两种方法主要通过分析词串内部词语之间的关系,来确定该词串是否是一个结构稳定的短语;而*N*-gram方法是结合词串所在的上下文信息,通过外部知识来判断该词串是否为一个结构完整的短语,文献[8]中的方法是基于这一设想。文献[9]中在抽取二元词汇搭配上将这几种计算方法做了比较。文献[10]中将C值与互信息进行结合进行术语抽取取得了较好的效果。本文使用基于统计的方法进行抽取,相关统计参数在二元算法的基础上进行了扩展,用以对多词短语的抽取。本文将短语的抽取看作是一个排序问题,选择互信息、C-value、似然比三种算法进行抽取,而后对结果集进行排序集成。互信息与似然比方法主要考察的是短语的内部结合度,而C-value考察的是上下文信息并且将词串长度加入到了考察范围。因此对这三种基础抽取方法进行集成,很好地融合了它们各自的优点,将短语的上下文、内部结合度及词串长度融为一体。

## 3 相关抽取方法

### 3.1 基于C-value的方法

C-value算法从根本上说还是基于频率的思想。以频率函数来衡量候选词串,通过这个词串在较长候选词串中的出现频率以及这些较长的候选词串数来确定候选词串是短语的可能性。但它参考了短语的长度和嵌套词的影响。它认为长度愈长的短语更难以出现,对于比较长的候选短语在其频率上应该有相应的加权。因为一些候选短语是被嵌套的词串,这样它的嵌套词会多次累计频率,所以需要进行相应的罚分来得到最终的分数。算法有三个方面的因子:(1)提取频率更高的词串;(2)对于更长候选词串的嵌入子词串进行罚分;(3)考虑候选词串的长度。具体的计算公式如下:

$$C-value(a)=\begin{cases}f(a) & (1)\\ f(a)-\dfrac{t(a)}{c(a)} & (2)\end{cases}$$

其中 $a$ 是候选词串,$f(a)$ 表示 $a$ 在语料库中出现的频率,$t(a)$ 是所有包含 $a$ 的较长候选词串出现的总次数,$c(a)$ 表示所有包含 $a$ 的候选词串的总数目。如果 $a$ 是最大长度的词串,则 $a$ 不被任何其他候选词串包含,此时候选串 $a$ 的唯一参数就是它们在集合中的出现频率,由式(1)计算得出。如果 $a$ 不是最长的候选词串,则有候选词串包括 $a$,则由式(2)计算。

### 3.2 互信息的方法

互信息是信息论中的一个概念,它用来度量一个消息中两个信号之间的相互依赖程度。二元互信息[6]是两个事件的概率函数,设两个待识别的字串为 $x$ 和 $y$,则在信息论中两个事件的互信息计算如下公式:

$$MI(x,y)=\mathrm{lb}\frac{P(x,y)}{P(x)\times P(y)} \tag{3}$$

如果 $x$ 和 $y$ 在一起出现的机会多于它们随机出现的机会,那么 $P(x,y)>>P(x)\times P(y)$,即字符串 $x$ 和 $y$ 结合十分紧密,则依据公式(3)计算的字符串互信息就比较大;反之 $P(x)\times P(y)>>P(x,y)$,这样计算出来的互信息就比较小。因此,可以利用互信息计算一个字串的内部结合强度,互信息值越高,$x$ 和 $y$ 组成短语的可能性越大;互信息值越低,$x$ 和 $y$ 组成短语的可能性越小。

传统的互信息方法如式(3),只能计算两个词之间的内部结合强度。为了适应抽取长度大于2的词串,Silva和Lopes将式(3)改进为:

$$MI(w_1,w_2,\cdots,w_n)=\mathrm{lb}\left(\frac{P(w_1,w_2,\cdots,w_n)}{AvP}\right) \tag{4}$$

其中:

$$AvP=\frac{1}{n}\sum_{i=1}^{n-1}P(w_1,w_2,\cdots,w_i)P(w_{i+1},w_{i+2},\cdots,w_n),$$

$$n\geqslant 2,P(w_1,w_2,\cdots,w_n)$$

$n\geqslant 2$,$W=w_1,w_2,\cdots,w_n$ 是多字串在给定语料库中所出现的概率。对于概率 $P(w_1,w_2,\cdots,w_n)$ 不能直接计算,可以利用MLE方法估计得到,具体公式如下:

$$P(w_1,w_2,\cdots,w_n)=\frac{f(w_1,w_2,\cdots,w_n)}{N} \tag{5}$$

其中 $f(w_1, w_2, \cdots, w_n)$ 表示多字串 $W$ 在该语料库中所出现的频率。$N$ 表示该语料库中的总字数。

### 3.3 卡方检验

卡方检验是一种常用的假设检验的统计学方法，主要研究两个变量间的关联性及频数分布的拟合度。

假设 $H_0$ 表示词 $w_1$，$w_2$ 是完全独立产生的，则它们偶然在一起的概率可以表示为：$P(w_1w_2)=P(w_1)P(w_2)$。如果语料中共有 $N$ 词次，则 $X^2$ 统计量计算了观测值和期望值之间差别的总和，将期望值作为比例因子。$X^2$ 的计算公式如下：

$$X^2=\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \tag{6}$$

其中 $i$ 表示表1中行变量，$j$ 为列变量，$O_{ij}$ 表示单元 $(i,j)$ 的观测值，$E_{ij}$ 表示期望值。当数值很大时 $X^2$ 满足卡方分布，对比表1中的观测频度和期望频度以验证是否独立，如果它们之间的差别很大时，可以否定它们是独立的 $H_0$ 假设。

**表1　$w_1$ 和 $w_2$ 的依赖关系表**

| 频次 | $w_1$ 出现 | $w_1$ 不出现 |
|---|---|---|
| $w_2$ 出现 | $O_{11}$ | $O_{12}$ |
| $w_2$ 不出现 | $O_{21}$ | $O_{22}$ |

通过计算边缘分布可以得到期望频度 $E_{ij}$ 的值，对表1形式的统计表，计算公式如下：

$$X^2=\frac{N(O_{11}O_{22}-O_{12}O_{21})}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{12}+O_{22})} \tag{7}$$

当置信水平为0.05时，临界值 $X^2=3.841$，即只有当计算值小于3.841时，有95%的置信概率认为 $w_1w_2$ 不是一个短语。

### 3.4 似然比方法

似然比(log-likelihood ratio)最初是由Ted Dunning提出来的。它虽然是一个简单的比值，但可以表达出一个假设的可能性比其他假设大多少。对于稀疏数据，似然比比卡方检验更加合适，而且，计算出来的似然比统计值比卡方检验的统计值更有可解释性。用参考文献[5]的两个可选的假设来解释二元组 $w_1w_2$ 的出现频率。

**假设1**　$P(w_2|w_1)=P=P(w_2|\neg w_1)$

**假设2**　$P(w_2|w_1)=P\neq P(w_2|\neg w_1)$

假设1是独立性假设的形式化，即 $w_2$ 的出现和前面 $w_1$ 的出现是独立的；假设2是非独立性假设的形式化，即 $w_2$ 的出现和前面的 $w_1$ 的出现是相关的。

使用最大似然估计的方法计算 $P$、$P_1$ 和 $P_2$，用 $c_1$、$c_2$ 和 $c_{12}$ 来表示在语料库中 $w_1$、$w_2$ 和 $w_{12}$ 出现的次数，则其计算公式分别如下：

$$P=\frac{C_2}{N} \tag{8}$$

$$P_1=\frac{C_{12}}{C_1} \tag{9}$$

$$P_2=\frac{C_{12}}{N-C_1} \tag{10}$$

假设二项式分布：

$$b(k;n,x)=\binom{n}{k}x^k(1-x)^{(n-k)} \tag{11}$$

实际观测到的 $w_1$、$w_2$ 和 $w_1w_2$ 频率的似然值是：

$$L(H_1)=b(C_{12};C_1,P)b(C_2-C_{12};N-C_1,P) \tag{12}$$

(假设1的情况)

$$L(H_2)=b(C_{12};C_1,P_1)b(C_2-C_{12};N-C_1,P_2) \tag{13}$$

(假设2的情况)

似然比 $\lambda$ 的对数值如下：

$$\begin{aligned}\mathrm{lb}\,\lambda=\mathrm{lb}\frac{L(H_1)}{L(H_2)}=\mathrm{lb}\frac{b(c_{12};c_1,p)b(c_2-c_{12};N-c_1,p)}{b(c_{12};c_1,p_1)b(c_2-c_{12};N-c_1,p_2)}=\\ \mathrm{lb}\,L(c_{12};c_1,p)+\mathrm{lb}\,L(c_2-c_{12};N-c_1,p)-\\ \mathrm{lb}\,L(c_{12};c_1,p_1)-\mathrm{lb}\,L(c_2-c_{12};N-c_1,p_2)\end{aligned} \tag{14}$$

其中，$L(k,n,x)=x^k(1-x)^{(n-k)}$。

使用似然比检验的优点在于：一是它有一个很清晰直观的解释，即如果似然比很小，表示它非常可能符合假设2，即 $w_1w_2$ 不是偶然出现的。二是它比卡方检验更好地解决了稀疏数据问题。这是检验两词串的有效方法，但是对于多词串却无法使用。为了适合多词串的似然比计算将公式从新定义[8]如下：

$$\begin{aligned}\mathrm{lb}\,L(w_1,w_2,\cdots,w_n)=2(\mathrm{lb}\,L(\frac{kf_1}{nf_1},kf_1,nf_1)+\mathrm{lb}\,L(\frac{kf_2}{nf_2},kf_2,nf_2)-\\ \mathrm{lb}\,L(\frac{kf_1+kf_2}{nf_1+nf_2},kf_1,nf_1)-\\ \mathrm{lb}\,L(\frac{kf_1+kf_2}{nf_1+nf_2},kf_2,nf_2))\end{aligned} \tag{15}$$

其中，

$$kf_1=f(w_1,w_2,\cdots,w_n),\ kf_2=AvY-kf_1,$$

$$nf_1=AvX,\ nf_2=N-nf_1$$

$$AvX=\frac{1}{n-1}\cdot\sum_{i=1}^{n-1}f(w_1,w_2,\cdots,w_i)$$

$$AvY=\frac{1}{n-1}\cdot\sum_{i=2}^{n}f(w_1,w_2,\cdots,w_i)$$

## 4 排序集成方法

排序集成的方法已经被广泛研究和应用[11]，但是将它应用到短语抽取上还不多。这里首先引入排序集成中的几个概念。

**定义1**(K-distance)

$L_1$ 和 $L_2$ 是基于同一候选集合 $(1,2,\cdots,n)$ 的两个排序，对于任意两个候选项 $i,j\in(1,2,\cdots,n)$，如果有 $L_1(i)<L_1(j)$ 且 $L_2(i)>L_2(j)$，则它们构成一个逆序对。K-distance$(L_1,L_2)$ 就是这两个排序的所有逆序对的个数。

**定义** 2(孔多赛标准)

将每一个候选项与其他选项一一对比,如果一个候选项在大多数投票上的得分高于另一个选项,那么它便击败了那个选项,击败所有其他候选项的便是孔多赛赢家。这种方法被称为孔多赛标准。

**定义** 3(Kemeny最优)

有 $m$ 个已经生成的排序序列 $(L_1, L_2, \cdots, L_m)$,序列 $L$ 是根据这 $m$ 个序列的重排序,如果 $L$ 使得 $S_k(L, L_1, L_2, \cdots, L_m)$ 达到最小值,那么 $L$ 为序列集 $(L_1, L_2, \cdots, L_m)$ 的Kemeny最优。其中,

$$S_k(L, L_1, L_2, \cdots, L_m) = \sum_{i=1}^{m} K - distance(L, L_i) \tag{16}$$

Kemeny最优符合孔多赛标准,但是当序列个数大于3个时,Kemeny最优就是一个NP难问题。因而Cynthia Dwork等人在元搜索引擎的开发时提出局部Kemeny最优的概念。

局部Kemeny最优:如果任意转换一对相邻候选项的位置,不存在序列 $Q$ 使得 $S_k(Q, L_1, L_2, \cdots, L_m) < S_k(L, L_1, L_2, \cdots, L_m)$,那么序列 $L$ 是序列集 $(L_1, L_2, \cdots, L_m)$ 的局部Kemeny最优。

基础集成方法:

波达计数[11]是一种投票机制方法。目前的投票方法有两种:一是多数决策;另一个是加权决策[12]。波达计数是多数决策,文献[13]中使用基于加权决策投票的方法对术语进行了抽取。各个统计抽取算法根据自己的判别标准对于各个候选词串进行抽取排序。如果候选者在选票中排第一位,它就得最高分值;排第二位得一个稍小的分值……依此类推。通过候选词串在序列中的位置来确定分值,最后的投票积分之和越高,说明该候选词串的表现越好。设 $t$ 为一个抽取算法所产生的候选词串序列,如果候选词串 $i \in t$,则 $t(i)$ 表示候选词串 $i$ 在 $t$ 中的位置。计分公式为:

$$w_t(i) = 1 - \frac{t(i) - 1}{|t|} \tag{17}$$

其中 $t(i)$ 为候选词串在排序中的位置,$|t|$ 为候选词串序列的长度。

除波达计数外常用的还有均值,几何均值等基础集成排序。顾名思义,均值是计算候选项在不同排序集中的排名均值,而几何均值是计算排名的几何均值。

Kicker方法[11]是在波达计数的基础上的改进。该算法需要记录候选词串 $i$ 在序列 $t$ 的前 $n$ 项中出现的总次数 $c(i)$。候选词串 $i$ 遍历所有的序列。如果 $i$ 在 $t$ 的前 $n$ 项中出现过,则 $c(i)$ 加1,若没有则扫描下一个序列,直到所有的序列都进行了扫描。计分表达式为:

$$Kicker(i) = w_t(i) \cdot \mathrm{lb}(c(i)) \tag{18}$$

其中 $w_t(i)$ 为波达计数如公式(14)所描述。Kicker方法是在波达计数的基础上,增加了对于候选词串在单个序列 $t$ 中的衡量。波达计数是对于候选词串整体分布的评估,而每个独立的抽取算法代表一个独有判别标准。这里的 $c(i)$ 可以看作一个信用评级,如果 $i$ 在一个抽取算法产生的序列的前 $n$ 项中出现,则 $c(i)$ 的评级加1。若候选词串 $i$ 在越多的序列中出现,$c(i)$ 的值越大,则表明 $i$ 被越多的算法信任,$i$ 成为固定短语的可能性就越大。

本文中的集成算法是先由各单独抽取算法进行抽取排序形成排序集,而后使用基础集成方法进行集成,最后使用局部Kemeny最优化算法来确定最后的抽取序列。文献[15]对七种单独抽取算法进行了集成,这些基础的抽取方法着重考察的不是短语的上下文信息就是短语的内部结构,因此集成投票实际上是短语的上下文与内部结构两种信息在投票。过多的基础抽取方法存在对上述两种信息的重复,如果方法组合选择不当还会造成不公平。

## 5 抽取算法

在文献[15]中使用了先计算二词串的各个统计参数,然后将符合约束条件的二词串定为种子,然后由种子向前和向后依次扩展一个词,计算此扩展词串的统计参数,如果符合约束条件则定为新的种子,直到设置的词串长度 $L$ 为止。此算法需要多次遍历整个语料,进行切分以及参数的计算,这是许多相似算法的一个弊端.另外本文是基于排序集成方法进行抽取故而每个单独的抽取算法都需要相同的前期处理。本文设计了一个新的整体抽取方法,其主要思想:一是根据种子长度分组并按分组依次计算种子的统计信息,分组处理降低了算法对内存的要求使该算法适用于处理大规模语料而且因为有分组的存在可以按分组搜索,提高了搜索效率。二是一次性计算此种子的所有抽取算法值并根据各个阈值对种子进行删减。每一个单独抽取算法所需的计算参数大致相同,计算一个抽取算法值的同时这些参数也可以被其他抽取算法使用,一次性方法减少了搜索语料的次数,从而提高了算法的效率。

抽取算法主要有三个阶段,首先确定种子,然后对不符合条件的种子进行删减,最后就是判断哪些是固定短语。下面将详细介绍这三个阶段。

### 5.1 确定种子

**步骤** 1 读入语料库 $B$。

**步骤** 2 利用标点符号等信息将句子粗分为较短的子句,而后对子句进行以词为单位的全切分,并按照切分出来的词串长度分别放入不同的文件中。这里将这些词串定义为种子。

**步骤** 3 对切分出来的文件进行统计形成数据字典文件,包括种子出现的次数、频率等信息。

### 5.2　删减种子

**步骤1**　利用数据文件中种子的频次，频率信息，首先计算长度为2的种子文件中所有种子的统计参数，如果某一个种子的参数值不在阈值范围内则将它删除，并记录在删除列表delete_list中，称其为非种子。

**步骤2**　依次计算长度为3，4…直至 $N$ 的种子文件中的种子。如果种子中含有delete_list中的非种子词串，则将其删除，如果不含非种子词串，则计算其参数值，并按照第一步中的方法判断是否将它移入删除列表。

### 5.3　短语的判定

将长度大于等于2的所有剩余的种子合并到一个节点序列中（这里的节点包括种子词串、词串长度、频率值（FT）、C-value（CV）、互信息值（MI）、似然比值（LR）），根据下列条件进行固定短语的判断：

（1）如果种子 $a$ 是种子 $b$ 的子词串，有相同频率并且长度相差为1，则 $a$ 不是固定词组。

（2）将符合标准的种子分别按照FT、CV、MI、LR降序排列，本文中不再单独生成排序序列而改用在种子节点中记录其在这种排序中的排序位置，即分别将IDFT、IDCV、IDMI、IDLR写入节点中。

（3）按照排序集成的原理对种子在四种排序中的位置进行综合计分，并依此分值从新排序，再使用局部Kemeny最优化方法求得最优排序，在这个排序集中靠前的种子就是要抽取的固定短语。下面介绍计分方法。

在短语抽取的过程中发现越是长度大的词串出现的频率就越低，在排序中越靠后，也就容易被漏掉。为照顾长词串，本文设计了一个新的计分方法，公式如下：

$$K(i)=t(i)\times \mathrm{lb}(c(i)) \tag{19}$$

其中 $i$ 表示候选词串，$t(i)=\sum_{j=1}^{m}\frac{1}{ID_{L_j}}$，$c(i)$ 表示 $i$ 的词串长度；$L_j\subset R$，$R$ 为排序集 $R(L_j\cdots L_m)$；$ID_{L_j}$ 为 $i$ 在序列 $L_j$ 中的排序号。

## 6　实验结果及分析

### 6.1　测试语料库

所用的语料库为2008年1月31天的新疆日报语料库，该语料库是已经过词附加成分切分及词性标注的XML格式，包含646篇文章，共31 695条语句，本文主要使用其词干信息。

### 6.2　实验结果

为评估排序集成方法的有效性，本文首先对互信息、C-value、似然比方法进行了参照实验，将抽取结果作为对比的基础。本文集成方法共得到候选短语4 023个，全面准确率为77.10%，比单独用互信息方法的52%准确率有提高，比C-value的平均准确率54.09%也改善了很多。前1 000个短语的准确率达到了86.0%。前 $K$ 个词（$K$ 取值100，500，2 000）正确率与直接抽取算法的对比如表2所示。

**表2　准确率对比**　（%）

| 抽取方法 | 互信息 | C-value | 似然比 | 排序集成 |
|---|---|---|---|---|
| 前100词 | 87.9 | 88.7 | 89.4 | 93.1 |
| 前500词 | 77.1 | 72.3 | 79.6 | 87.5 |
| 前2 000词 | 68.4 | 65.6 | 66.3 | 79.5 |

与文献[14]中所用集成方法的前2 000词的72%准确率相比，本文算法的准确率也有提高。在所抽取的4 023个短语中，对不同长度词串的抽取准确率做了一个统计。详细数据如表3。

**表3　不同长度词串的准确率对比**

| $N$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 准确率/(%) | 80.4 | 78.0 | 72.0 | 70.6 | 78.8 |

### 6.3　结果分析

由实验数据可以看出排序集成方法是有效的。它很好地整合了三种抽取算法的特点，既有C-value对词串上下文信息的考虑，又有互信息、似然比对词串内部结合度的考察。本文设计了一个整体的短语抽取方法，可以一次性得到三种抽取方法的短语及其在每种方法中的排序信息，相对于文献[14]中分别使用单独的方法进行抽取再进行集成，在算法效率上有很大提高。文献[15]中使用种子扩展的方法，一步一步将种子扩展到术语长度，本文中设计了一个种子删减的算法，一次生成所有的种子，而后对不符合的进行删除。该方法省去了多次对语料的切分也提高了结果的准确率。但是高的准确率是在种子删减过程中使用了严格的删减制度产生的，即如果种子有一个抽取算法值不满足阈值要求则将它删除。长词串的正确率有很大提高，说明在基础集成算法中加入词串长度起到了一定作用。哈萨克语是一种形态丰富的语言，每个词在不同的上下文中都有不同的变化形式，如果将每一种变化形式都认为是单独的词必将导致严重的数据稀疏，而词干是一个词中体现词汇意义的部分，故本文选择词干作为词的代表进行统计，实验结果表明选择是正确的。本文的方法主要是基于统计学的，除了前期针对哈语的特点而做的语料预处理，其他的算法完全适用于其他语言。

## 7　结论

本文采用排序集成的方法将C-value、互信息和log-likelihood三种统计方法有机融合在一起，提高了抽取的正确率。本文抽取结果基本达到了预期，但是还有很大的提升空间，集成方法的研究将是接下来的工作重点。努力减少算法的时间、空间等复杂度，使得集成算法能够胜任大数据量、更多统计参数的集成工作。

（下转223页）