Research on Chinese N-gram Statistical Rule and its Application

Zhaoming Yin[1], Huarui Zhang[2]

1. School of Software and Microelectronics, Peking University, Beijing, China

2. Institute of Computational Linguistics, Peking University, Beijing, China

1. stplaydog@gmail.com, 2. hrzhang@pku.edu.cn

Abstract: In this article, we assign Chinese n-gram sequences to different types by their statistical properties such as frequency, mutual information and left/right border entropy. We call these sequence type "Radixes" and define some combination rules between them. Based on the radixes we classified and their combination rule we designed a new Chinese segmentation algorithm without dictionary based on dynamic programming, and do some research on the automatic word extraction of Chinese words consist of 2 to 4 letters, we achieved good performance on some aspects.

Keywords: Artificial Intelligence; Chinese word segmentation with no dictionary; dynamic programming, new word extraction

中文 n 元字串统计规则及其应用研究

尹朝明[1]，张化瑞[2]

1、北京大学软件与微电子学院， 北京，100089

2、北京大学计算语言研究所， 北京，100089

1. stplaydog@gmail.com, 2. hrzhang@pku.edu.cn

【摘要】：本文通过观察汉字 n 元字串在表叙其作为单位个体的频率、互信息、左右熵等统计量的特征，总结出了几种不同特点的中文统计串类型，我们称之为“基”，并人工规定了不同的“基”之间的组合规则。我们在“基”和组合规则的基础上提出了基于动态规划的无词典分词算法，并进行 2 至 4 字词的词汇提取实验，取得了较好的效果。

【关键词】：人工智能；无词典分词；动态规划；新词识别

1 引言

随着 web2.0 技术的不断发展，语言在这一更加自由开放的信息载体的滋养下不断的发展，并衍生出了许多新词或者新的表达方式和意义。中文 n 元字串的统计信息如频率、互信息、字串边界左右熵等不依赖于词典，句法信息，反映了中文字与字之间在统计上的结合程度，和其作为单独的逻辑块的自由使用程度，因而在中文未登陆词的识别中得到了广泛研究。清华大学的罗盛芬，孙茂松[1][3]在选取基于 n 元字串的统计量，及这些统计量在识别二字词的性能进行了比较深入的研究，北京大学的谌贻荣[2]组合了几种不同的统计量并提取了“单元度”概念，其实验计算双元素构成的片断时比常见算法中最好的互信息在精确率和召回率相等时的 F-measure 值高出 26.8%，将二字词提取的能力提高到了新的层次。本文在前人研究的基础上，提出了将统计串按照其在统计量上的不同特点归类于不同的汉字统计“基”的概念，并将这些基应用到了无词典分词之中，并将切出来的词作为识别出来的词，通过循环切词的方法，我们进一步可以获得长词或者结合到短语提取的工作中[4]。

2 统计规则的基础概念

定义 1：单元基（Unit Radix，简称 UR），表示词或短语的有意义的基本组成单位如“孙中山故居”中的“孙中山”等。

定义 2：左连接基（Left Radix，简称 LR），表示单独不能成单元基，但是有非常大的概率与其右部出现的字串单位结合成单元基，如“阿拉”之于“阿拉伯”就是一个左连接基。

定义 3：右连接基（Right Radix，简称 RR），表示单独不能成单元基，但是有非常大的概率与其左部出现的字串单位结合成单元基，如“斯基”之于“诺维斯基”就是一个右连接基。

定义 4：双连接基（Double Radix，简称 DR），表示单独不能成单元基，但是有非常大的概率与其左部和右部出现的字串单位共同结合成单元基，如“鲁木”之于乌鲁木齐。

定义 5：连接基（Connective Radix，简称 CR），在这里我们统一称左连接基，右连接基，双连接基为连接基。

定义 6：单字基（Single Radix，简称 SR），在汉语中大部分的字是单元基，其本身就是语言学中的“语素”，如“我”等，有少部分的字是连接基，如“琵”之于“琵琶”等，我们将字定义为单字基。

基与语言中的词或短语的区别在于其是一个递归，可发散的概念，其既可以是最基本的语素，又可以是短语，甚至可以是由短语组成的短语。一个基可以由其他的基组合而来，同时组合这个基的元素亦由其他的基所组成。基的组合规则如表 1 所示：

978-1-935068-01-3

Table 1. Combination rule of Radix
表 1. 基与基之间的组合关系

规则序号	规则表示	范例	
1	$\sum_{i=1}^{n} {}^{0}SR_i = {}^{1}UR^n (n \geq 2)$	n=2	SR（葡）+SR（萄）=UR(葡萄)
		n=3	SR(真)+SR(善)+SR(美)=UR(真善美)
2	$\sum_{i=1}^{n} {}^{0}SR_i = {}^{1}CR^n (n \geq 2)$	n=2	SR(真)+SR(善)+SR(美)=CR(真善美)
3	$\sum_{i=1}^{n} {}^{k-1}UR_i = {}^{k}UR^n \ (n \geq 2)$	n=3	SR(多)+SR(快)+SR(好)=CR(多快好)
		n=2	UR(孙中山)+UR(故居)=UR(孙中山故居)
		n=3	UR(静静)+UR(的)+UR(顿河)=UR（静静的顿河）
4	$LR+(RR \vee SR) = {}^{k}UR^2$	LR+RR	LR(布达)+RR(佩斯)=UR(布达佩斯)
5	$(SR \vee LR)+RR = {}^{k}UR^2$	LR+SR	LR(阿根)+SR(廷)=UR(阿根廷)
		LR+RR	LR(布达)+RR(佩斯)=UR(布达佩斯)
		SR+SR	SR(毛)+RR(泽东)=UR(毛泽东)
6	$LR + SR = LR^2$	SR+SR	LR（乌鲁)+SR(木)=LR(乌鲁木)
7	$SR + RR = RR^2$	SR+ RR	SR(生)+RR(梦死)=RR(生梦死)

我们选取 n 元字串的如下几个统计量来描叙成基的统计特征，分别是字串的频率，互信息和左右熵。在判定基的打分公式中，我们假定频率为阈值属性，左右熵为成基属性，互信息为基本属性，既一个基的得分由其互信息决定为互信息乘以一个参数 C，而 C 的值由阈值属性频率决定如当频率大于 10 时为 3.5 小于 10 时为 3.0, 我们在实验中采用的就是此参数，一个字串归于哪个基由成基属性决定，左右熵都不为零为自由基，如当左熵等于零时其为右基等。

3 统计规则在自然语言处理中的应用

基于统计规则的无词典分词方法是在动态规划算法的基础上建立的，因为我们已经得到了每个基的得分，并且人工设定了基与基之间的组合规则，因此我们就可以利用基的得分和组合规则从一个字串的侯选搜索空间中找到得分最高的切分方式作为切词的结果。

我们在动态规划的算法中使用如下的递归公式，其中 $f(0,n)$ 代表长度为 n 的子串的最佳得分，i 为切分此串的位置，p 为子串 $f(i+1,n)$ 对应的规则权值。

$$f(0,n) = \max_{i=0}^{n-1}(f(0,i) + pf(i+1,n))$$

3.1 无词典分词中的统计规则和语言规则

我们在分词的过程中为在某个统计规则和语言规则模式下的串的得分乘以一定的权值，表 2 是我们总结的不完全的统计规则和语言规则，对应的规则权值矩阵不在此列出。

Table 2. Statistical and linguistic rules in segmentation
表 2：切词中使用到的统计和语言规则集合

统计规则	
加分规则	LR+RR 的规则为加分规则
减分规则	不在规则集合中的规则，如 RR +LR 等
其他规则	规则集合中其他规则不加分也不减分
语言规则	
独立性规则	大多数时候是以单独的有意义的字的形式出现的字，如：“的”、“地”。
粘连性规则	由于语料的切分规范，有些词被粘连在一起，比如“一九九八年”数词和年被粘连，此规则被添加到了规则集中。
符号性规则	标点符号是作为单独的个体出现的，为了避免切分中将标点符号与词粘连，我们将符号性规则放入规则集合中。

3.2 无词典切词算法

由以上的递归公式和规则权值矩阵，我们给出如下的基于动态规划的无词典切词算法，算法所用到的数据结构和算法框架分别如表 3 至表 4 所示。

算法说明：1：为了便于处理，我们假设一个基只与前一个基之间存在规则关联，而与历史出现的基是独立的。2：由于提取出来的基仅仅按照得分的高低进行排列，我们并没有绝对的基与基之间的规则，我们的规则就是为某两个基之间的组合赋予对应的权值。3：动态规划算法本身的复杂度为 $o(n^2)$，但是由于中文最长的词是有一定限制的，我们可以修改算法，使其只对前 max_len 个子串进行计算避免了无谓的匹配，而这样的算法的复杂度与正向最大匹配算法的复杂度是一致的为 $o(m \times n)$,其中 n 为字串长度，m 为中文最长词长。4：为了将问题简单化，初始分词的基的度统一为 2，我们可以通过迭代的方法得到长词的基，显然，当词的长度大于 2 的时候其基的级大于等于 2，我们将这些新识别的基作为后续的新词识别的内容加以讨论。

3.3 词语识别

新词识别是与无词典分词共同进行的，我们选取基中成词置信度（得分）较高并且在切词阶段在上下文中能够独立分开的部分作为识别的新词，并且用低层级的基切词的结果迭代进一步获得高层级的基并作为下一轮切词的基础。在得分的最后乘上了一个长度权值，给层级高的基

以更高的得分给词按照不同的长度以不同的权值。

4 实验及结果分析

我们选取人民日报 1998 年 1 月语料去除标记、分词信息，第一轮计算度为 2 的基，继续计算第一轮切词后的统计属性，得到级为 2 的基与第一轮的基合并，依此类推。我们定义两种不同的判断切词效果的标准，第一种标准是传统的准确率-召回率-F 值策略，另一种是计算程序切分的句子与正确切分句子之间的词条编辑距离作为切分的错误数，进而计算准确率-召回率-F 值。切词实验和词汇提取的结果分别如表 5-6 所示：

Table 3. Data structure of basis_set, rule_set and track
表 3：数据结构 basis_set, rule_set, track 的相关说明

basis_set	rule_set	track

• 成员：Map: <word, feature>
• Feature 成员：
Type_num: basis[]
Score_num: score[]
• 函数：#返回一个字串可能的基的组合：getbasis(word)
#返回一个字串可能的基的组合对应的得分：getscore (word)
	• 成员：Map: <word, power>
	Map: <(basis1,basis2,power>
	• 函数：#返回基与基之间的组合所对应的打分权值：
	sta_rule(basis1, basis2)
	#返回某个子串所含有的语言规则的打分权值：
	sta_rule(basis1, basis2)
		• 成员：
		#记录此位置最高总得分
		Score
		#记录此位置的基的类型
		Basis
		#记录此位置的前一个最优位置
		pre_position
		#记录此位置对应字串 Word

Table 4. Dynamic programming algorithm based on statistical rule
表 4：基于统计规则的动态规划分词算法

DYNAMICSEGMENTATION(input, output, basis_set, rule_set, track)
#Input 是待分词的字串，output 是分好词的字串，
• 从 input 中的由第一个字到第 i 字所组成的子串：
• 初始化此位置的得分为 track[i].score = track[i-1].score
• 在此子串的第一个切分位置到第 i-1 个切分位置 j：
• former_basis = track[j].basis • former_score = track[j].score
• 对于从第 j 个位置切出来的子串 subStr(j,i)的第 k 一个可能的基：
• this_basis = basis_set.getbasis(input(j,i))[k]
• this_score = basis_set.getscore(input (j,i))[k]
• 若：
tmp_score=former_score+this_score*rule_set.sta_rule(former_basis,this_basis)*rule_set.lan_rule(subStr(j,i))①*levelPower >track[i].score
• track[i].score = tmp_score • track[i].basis = this_basis
• track[i].pre_position = j • track[i].word = input (j, i)
• 从 track 数组中的最后一个元素 m 开始，当 m 不等于 0 时：
• output = output + track[m] .word • m = track[m] .pre_position
• 返回 output

Table 5. Result of segmentation without dictionary
表 5：无词典切词结果

	第一轮	第二轮	第三轮
准确率	0.934	0.929	0.929
召回率	0.928	0.926	0.927
F 值	0.931	0.928	0.928
自定义准确率	0.696	0.696	0.695
自定义召回率	0.667	0.684	0.685
自定义 F 值	0.681	0.69	0.689

从切词实验数据可以得出，随着递归层数的递进，在传统的准确率-召回率-F 值的评判上，基本上都差不多，而在自定义的传统的准确率-召回率-F 值评判上，准确率都有小幅下降，而召回率有所上升，说明算法有倾向于长此的特性，这可以通过调整词长权值来修正。

从词汇提取的实验数据来看，长词词汇提取的结果不甚理想，主要原因有两点：

1）与语料切分的标准有关系，1998 年 1 月语料切分的粒度比较低如：“张大方”被切分成为“张 大方”。

Table 5. Result of word extraction grouped by word length
表 6：词汇提取结果及其随频率分布

频率	二字词			三字词			四字词		
	实际数	切分数	正确数	实际数	切分数	正确数	实际数	切分数	正确数
1	12361	5649	2255	4979	13364	1126	2839	8850	537
2	4207	2335	1482	1436	2117	438	854	943	223
3	2417	1304	1150	675	818	242	453	390	146
4	1566	919	852	484	459	166	257	165	85
5	1140	683	673	296	271	95	157	100	51
>5	8363	7047	7016	1397	1279	644	627	386	204
总计	30054	17937	13428	9267	18308	2711	5187	10834	1246

978-1-935068-01-3 © 2009 SciRes.

Figure 1. 2-letter word's precision recall, f-measure
图 1. 二字词随频率正确率召回率 F-值变化趋势

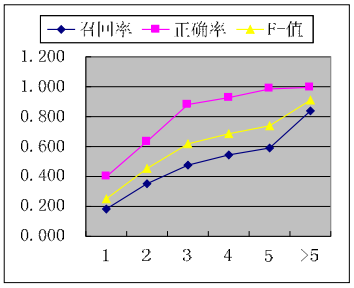


Figure 2. 3-letter word's precision recall, f-measure
图 2. 三字词随频率正确率召回率 F-值变化趋势

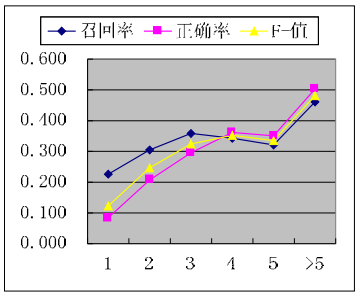
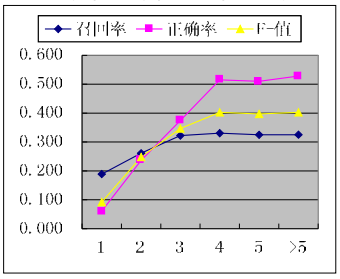


Figure 3. 4-letter word's precision recall, f-measure
图 3. 四字词随频率正确率召回率 F-值变化趋势



2）低频词所占比例较大，由于数据的稀疏使得低频词其频率不一定能够反映其真实特性，使得成基判断时发生失误。从图 1-3 可以看出，对于切分出来的词当其频率较高时可信度是比较大的，比如说二字词中频率大于 5 时的正确率能够达到 99.5%以上，这是我们已知的最好结果之一，同时在处理低频词特别是频率为 1 的词的时候切词算法也没有将其全部漏掉，说明纯粹靠属性打分方式无法提取的词，也能够结合上下文的语境通过切词来提取。

5 结语

本文提出了一种在中文 n 元字串基础上发展而来的字串成基规则，对字串结合的规则进行了初步的探讨，并在成基规则的基础上发散出了一种新的切词方法，在高频词的提取中取得了比较理想的结果，但是总体来说低频词的处理还有很大的提升空间，也是我们进一步的研究方向。

致谢

本文的书写，要感谢一直支持我鼓励我的俞敬松老师，在程序编写上给与我指导的李吉同学，谭大伟同学，同时，为了更进一步的在大数据集合上进行我们深入的研究，还要感谢李泉同学无私的提供了相关的大规模语料库，最后感谢在研究路上提供无私资助的父母，和北京大学校团委学术科创部的挑战杯课外研究基金。

References（参考文献）

[1] Shengfen Luo and Maosong Sun. Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures[A]. In: *Proceedings of Second SIGHAN Workshop on Chinese Language Processing*[C]. Sapporo,Japan: 2003 24～30

[2] Smadja. F. Retrieving collocations from text : Xtract[J]. *Computational Linguistics*,1993, 19(1): 143 -177.

[3] Thomas H. Cormen, Charles E. Leiserson, et al. Introduction to Algorithms[M]. Boston: Addison Wesley, 1989

[4] Yirong Chen. An Algorithm of Unithood Based on Internal Connection Rate and Marginal Freedom Rate[A] In: *JSCL—2005*, 315～318
谌贻荣. 内部紧密度和边缘自由度相结合的符号串单元度计算[A]. 见:自然语言理解与大规模内容计算[C]、北京：清华大学出版社，2005 315～318

[5] Shengfen Luo and Maosong Sun. Chinese Word Extraction Based on the Internal Associative Strength of Character Strings[J] Journal of Chinese Chinese Information Proceeding, 2003, 17(3): 9-14 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

978-1-935068-01-3 © 2009 SciRes.