

基于规则剔除和统计的分词词表选词方法

刘小勤 郭炳炎 刘开瑛

山西大学计算机科学系 030006

摘要: 本文介绍了一种规则剔除和统计相结合的分词词表选词方法。首先从生语料中取出所有的 N 元组 —— 相邻接的 N 个汉字串, 计算其串频、互信息和相关度等统计量, 然后采用规则集剔除噪声, 接着用三个统计量共同作用, 根据预定的召回率选出最小子表, 最后人工判断其是否为词, 并将非专名的未登录词加入基本词表中。从 412 万汉字的语料库中选出了 1598 个未登录词, 加入基本词表中, 形成了含词 79454 条的分词词表。

关键词: 串频 互信息 相关度 绝对功能字 相对功能字 功能词 规则

An Approach to Word Extraction for Segmentation Vocabulary Using Rule Elimination and Statistics

Liu XiaoQin Guo BingYan Liu KaiYing
Department of Computer Science, ShanXi University, 030006

Abstract In this paper, an approach based on rule elimination and statistics is developed to extract words for segmentation vocabulary. Firstly, every N-gram (a string of n adjacent characters) is extracted from raw corpora, then statistical items, including frequency, mutual information and coincidence degree, are calculated; secondly, the rules are used to eliminate noise; thirdly, n-grams in several minimal sub-tables which are extracted according to their frequency, mutual information and coincidence degree are regarded as potential unknown words; Finally, they are further categorized manually and those true unknown words are accepted except proper noun. We build a corpus, which contains about 4,120,000 Chinese characters, to test the effectiveness of this method. 1598 unknown words are added to basic vocabulary and the final segmentation vocabulary contains 79454 words.

Keywords Frequency, Mutual Information, Coincidence Degree, Absolute Function Character, Relative Function Character, Function Word Rules

一、介绍

随着中文信息处理各个应用领域的发展, 分词词表的研制已越来越引起人们的重视。

语言学词典是分词词表的一个重要来源, 但又缺少一个完整的实用的分词词表中必不可少的几部分词, 如“见字明义”的词, 新词以及比较通用的领域词汇等。因此, 我们汇集了《现代汉语词典》^[1]、《现代汉语规范词典》(审核稿)^[4]、《信息处理用现代汉语常用词表》^[5]等共三部词典, 通过统一筛选从中得到 77856 条不同的词条, 作为分词词表的

基本部分，称之为基本词表。称不在基本词表中的词为未登录词。对于未登录词，采用基于语料库的半自动的选取方法。具体做法是：

选取一定规模的生语料，从语料中抽取 N 元组（N 的取值为 1 到 4）并计算其串频、互信息、相关度等统计信息。对得到的 N 元组首先进行规则剔除，对规则剔除后剩余的 N 元组，用串频分段，生成若干一定串频范围内的全表，根据串频值选择不同的召回率，选出各自的最小子表。由于上述方法只能减少而不能完全消除噪声，所以对选出的最小子表人工判断，找出其中非专名的未登录词加入基本词表中，生成完整的带有多种统计信息的分词词表。

二、统计量的计算

(1) 串频

串频是指一个 N 元组 W 在语料库中的出现次数。记作 $f(W)$ 。

一个 N 元组是指语料中连续出现的 N 个汉字组成的一个字串。记作 W。例如 N=1 即为一元组，N=4 即为四元组。

(2) 互信息^[6]：

所谓互信息是一个信息论的概念，用来描述字串的组成部分之间结合的紧密程度。

一个 N 元组 W 的互信息 $mi(W)$ （以二元组为例）由下式定义：

$$mi(W) = \log_2 \frac{N_c^2 \times f(W)}{N_w \times f(A) \times f(B)} \quad \dots\dots (1)$$

其中 W 的首字为 A，次字为 B， $f(A)$ ， $f(B)$ ， $f(W)$ 分别表示 A、B 和 W 在语料库中的出现次数，即其串频。 N_c 为语料库中总字数， N_w 为语料库的总词数。

(3) 相关度^[7]：

在语料库中连续出现的汉字串，很可能构成一个中文词汇。也就是说，如果 N 元组能成词，则它的相邻词素之间就有较大的相关性。这种相关性可用 Pearson 的 χ^2 一统计量来度量，称之为相关度。

一个 N 元组 W 的相关度 $rel(W)$ （以二元组为例）由下式定义：

$$rel(W) = n \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - \frac{n_{i.} \times n_{.j}}{n})^2}{n_{i.} \times n_{.j}} \quad \dots\dots (2).$$

其中 W 的首字为 A，次字为 B，n 为语料库中二元组的总串频， $n = n_{11} + n_{12} + n_{21} + n_{22}$ 。 n_{11} 表示首字为 A，次字为 B 的串频、 n_{12} 表示首字为 A，次字非 B 的串频， n_{21} 表示首字非 A，次字为 B 的串频， n_{22} 表示首字非 A，次字非 B 的串频，记 $n_{i.} = n_{i1} + n_{i2}$ ， $n_{.j} = n_{1j} + n_{2j}$ ($i=1,2, j=1,2$)。

三、规则剔除

由于随意生成的 N 元组中不仅包含词且含有大量的噪声,有的噪声又很难通过统计量剔除。因此我们充分利用了较长的 N 元组包含信息量大的特点,对噪声采用规则剔除,根据规则所剔除的 N 元组数、误剔除的词数及与其他规则的交叉情况,提取规则构成规则集。其中四元组包含 8 条规则(R4-1—R4-8),三元组为 9 条(R3-1—R3-9),二元组为 7 条(R2-1—R2-7),共 24 条规则。为叙述方便,特作如下约定:

设四元组为 ABCD(A, B, C, D 分别为四个汉字),则称 AB, CD 都不是词的为 0 型四元组, AB 为词, CD 非词的为 1 型四元组, AB 非词, CD 为词的为 2 型四元组, AB, CD 都是词的为 3 型四元组, A=B, C=D 的为 AABB 型四元组。设三元组为 ABC(其中 A, B, C 分别为三个汉字),则称 AB, BC 都不是词的为 0 型三元组, AB 为词, BC 非词的为 1 型三元组, AB 非词, BC 为词的为 2 型三元组, AB, BC 都是词的为 3 型三元组。

以下为规则示例:

R4-2 若四元组中含绝对功能字,则四元组非词。

功能字(词)是指那些构词能力弱且在语料中出现次数较多的字(词),如“的”、“我”、“而且”、“什么”等。这些字大部分是副词、介词、连词等虚词以及代词。其中有的功能字一般只用作虚词或代词,如“的”、“我”、“最”等;而有的则兼作实词,构词能力仍比较强,如“边”、“单”、“自”等。称第一类为绝对功能字,第二类为相对功能字。

R3-3 若 0 型三元组含相对功能字,则该三元组非词。

R4-1 若四元组中含功能词且非 AABB 型,则四元组非词。

R2-7 若二元组串频等于 af,则二元组非词。

其中 af 为带前缀的 2 型三元组前两字或带后缀的 1 型三元组后两字与该二元组相同者的串频之和。

表 1 为上述 4 条规则的示例。从表中可以看出,这些噪声很难用串频、互信息和相关度的阈值将其剔除。另外经规则剔除后 N 元组的数量会大大减少,可显著减少其后的计算量。

四、实验结果与分析

实验用语料取自 1997 年中国人民大学报刊复印资料(光盘版),语料规模为 412 万汉字。对生语料抽取 1—4 元组,共形成一元组 4884 个,二元组 304971 个,三元组 973202 个,四元组 1449853 个。

表 2 为形成的 N 元组含基本词表中词条的情况。由表中可以看出,在形成的不同的二元组中词的比例相对较高,占总数的 8.43%,而不同的三元组和四元组中,分别只有 0.29%和 0.22%是词,也就是说,绝大部分 N 元组不是词。表中成词数指一定范围内的 N 元组含基本词表中的词数、成词率指成词数与该段 N 元组数之比(下同)。

表 1 规则剔除的 N 元组示例

规则	序号	N 元组	串频	互信息	相关度
R4-2 (带着重号者为绝对功能字)	1	他踉踉跄	1	5.02	3434102.50
	2	就意味着	40	-19.86	522382.92
	3	的基础上	390	-24.04	1110288.84
R3-3 (带着重号者为相对功能字)	4	趁田居	1	0.80	187456.89
	5	又不愿	15	-9.68	2455.64
	6	更好地	130	-8.38	54718.92
R4-1 (带着重号者为功能词)	7	敢惹咱俩	1	1.80	2289402.00
	8	人甚至耽	2	-15.84	1130932.44
	9	从根本上	113	-22.49	858576.65
R2-7 (带着重号者为前缀或后缀)	10	涂料	1	12.12	3467.74
	11	半闲	8	8.33	2365.51
	12	费者	531	6.40	43655.25

表 2 一到四元组成词情况表

	成词数	N 元组总数	成词率
一元组	4844	4884	99.18%
二元组	25698	304971	8.43%
三元组	2817	973202	0.29%
四元组	3252	1449853	0.22%

表 3 N 元组规则剔除统计表

	剔除的 N 元组数 (占 N 元组%)	误剔除的词数 (占成词数%)	误剔除率
二元组	170012 (55.75%)	394 (1.53%)	0.2317%
三元组	784492 (80.61%)	431 (15.30%)	0.0549%
四元组	1099878 (75.86%)	461 (14.18%)	0.0419%

表 3 为对 N 元组进行规则剔除的情况，其中误剔除率为误剔除的词数与剔除的 N 元组数之比。从表 2 和表 3 中可以看出，误剔除率大大低于所在 N 元组的成词率。例如四元组的成词率为 0.22%，其误剔除率为 0.0419%，在剔除了 75.86% 的四元组后，只剔除了 14.18% 的词。经过规则剔除后，N 元组的数量大大减少，而相比较而言，词的减少不是很多。可见这种剔除是非常有效的。

我们选取的三个统计量：串频、互信息、相关度均是 N 元组是否成词的很好的度量。见图 1 和图 2。其中图 1 为串频 \geq 给定值时，二元组的成词率。从图中可以看出，随着串频的增大，二元组是词的可能性亦大大增加了。如二元组串频 ≥ 1000 时，成词率高达 76.27%。图 2 为将全部二元组按互信息或相关度降序排序，平均分成十等份之后，每一

等份中的成词数。从图中可以看出，二元组的前几部分已经集中了绝大部分的成词词条。如二元组的前三等份成词数以相关度降序排序时为 17166 条，占二字的 66.80%，以互信息降序排序时为 14508 条，占二字的 56.46%。三元组和四元组亦有类似的情况。

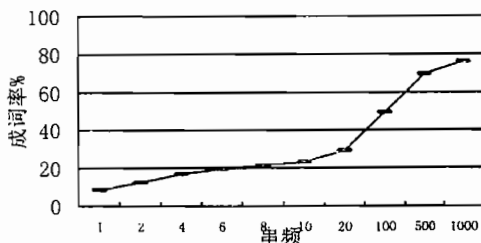


图 1 二元组串频与成词率关系示意图

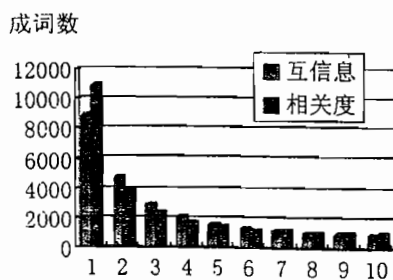


图 2 以互信息或相关度降序排序，每等份二元组中成词数

将规则剔除后的 N 元组分别按串频分成若干全表，任一全表按某种顺序排列后从全表的起始位置起到某个指定位置的部分构成该全表的一张子表，定义子表的召回率为子表的成词数与全表的成词数之比。在召回率一定时，我们希望子表尽量小，称此时的子表为最小子表。对串频低的全表则选择低召回率的最小子表，而串频高的全表则选择较高召回率的最小子表，最后对选出的最小子表人工进行检查，选出未登录词，并将非专名的未登录词加入基本词表中，形成最终的分词词表。

对最小子表的下列三种生成方法进行比较：

方法 1：最小子表中 N 元组互信息 \geq 阈值 T1

方法 2：最小子表中 N 元组相关度 \geq 阈值 T2

方法 3：最小子表中 N 元组互信息 \geq 阈值 T1 且相关度 \geq 阈值 T2

总的说来，方法 3 的效果最好。如在四元组串频为 1 的全表中，召回率为 20% 时，方法 1 的最小子表长度为 7741，方法 2 的为 11023。因此我们采用方法 3。

表 4 为对规则剔除后，采用方法 3，以串频分段，并取不同的召回率时的最小子表经人工检查后得到的未登录词（不含专名）的统计结果。从 2728026 个二到四元组中最终选出占总数 1.77% 的二到四元组共 48398 个，其中含基本词表中的词条 12142 条，占基本词表中二到四字词总数（共 31767 条）的 38.22%，另从中选出未登录词 1598 条加入词表中，形成了新的分词词表。表 5 为上面选出的未登录词示例。

五、小结

利用生语料抽取 N 元组可使我们摆脱分词的困扰，同时也会生成大量的噪声，这就使得选词变得更加困难。因此，我们利用规则剔除噪声之后，再用串频、互信息和相关度三者结合选词，实践表明，这种方法是切实可行的。需要指出的是，要研制一个实用的词表，语料规模需进一步加大。而且，从实验结果看，规则剔除非常有效但比较粗糙，且误剔除的词数也比较多，应进一步完善和扩大规则集。

表4 选词结果统计

	串频	召回率	T1	T2	最小子表	成词数	未登录词
二元组	1	20%	7.90	242.53	5506	1206	205
	2—3	30%	5.94	176.91	4275	1576	201
	4—9	40%	4.72	128.98	4215	2202	229
	10及以上	60%	3.64	229.04	7209	5108	220
三元组	1	20%	-4.11	2934.86	6743	146	103
	2	30%	-7.48	4722.99	2292	100	55
	3	40%	-8.60	538.12	1290	78	45
	4及以上	60%	-11.04	1337.77	6213	678	247
四元组	1	20%	-11.91	190860.45	6289	208	85
	2	30%	-17.00	508754.20	1115	157	34
	3	40%	-23.79	763136.99	402	106	6
	4及以上	60%	-27.85	39582.65	2849	577	168
小 计					48398	12142	1598

表5 未登录词示例

二元组	配送 吸纳 找到 抓好 餐饮 炒股 患儿 肺癌 恍若 胡侃 柜组 捧杀 尿检 骑警 灯箱 窃喜 坏帐 俏销 表姨 猫咪 丁亥 醉虾
三元组	桔红色 捣浆糊 老鼠会 批评家 枯燥性 娃娃家 偷漏税 服务器 姊妹篇 证券委 申请者 摆阔气
四元组	浮法玻璃 磕磕碰碰 假冒伪劣 基础设施 甩手掌柜 追涨杀跌 抓大放小 文化垃圾 原汁原味 狂轰滥炸 弯腰曲背 二律背反

参考文献

- [1] Jian-Yun Nie, Marie-Louise Hannan, etc. "Unknown Word Detection and Segmentation of Chinese using Statistical and Heuristic Knowledge", Communications of COLIPS, vol.5, No 1 & 2, DEC 1995, 69-77
- [2] Huang Xuan-jing, Wu Li-de et al., Statistical Acquisition of Terminology Dictionary, Fifth Workshop on Very large Corpora, Tsinghua University, 1997, 142-152
- [3] 中国社会科学院语言研究所词典编辑室编, 现代汉语词典, 商务印书馆, 1997.8
- [4] 李行健主编, 现代汉语规范词典(审核稿)(内部资料), 1998.
- [5] 刘源等, 信息处理用现代汉语分词规范及自动分词方法, 清华大学出版社等, 1994.6
- [6] 孙茂松等, 人机并存, “质”“量”合一, 语言文字应用, 97.1., pp.79-86
- [7] 黄萱菁等, 基于机器学习的无需人工编制词典的切词系统, 模式识别与人工智能, 1996.12., pp.297—30