动词与动词搭配方法的研究

白妙青[1]　郑家恒[2]
[1](山西大学计算中心,太原 030006)
[2](山西大学计算机系,太原 030006)
E-mail:bmq@sxu.edu.cn

摘　要　搭配是汉语自动句法分析的重要环节,而动词是句法分析的核心。论文面向中文信息处理,通过对真实文本的统计分析归纳了搭配自动获取规则,结合统计模型,研究了动词-动词搭配中各关系类型的分布特征以及搭配词语的位置分布特征,在此基础上成功地抽取出所选语料中的动词—动词搭配。其中抽取正确率为 75%,召回率为 64%。

关键词　词语搭配　语料库　中文信息处理

文章编号 1002-8331-(2004)27-0070-03　文献标识码 A　中图分类号 TP391

Study on Ways of Verb-Verb Collocation

Bai Miaoqing[1]　Zheng Jiaheng[2]
[1](Computer Center of Shanxi University,Taiyuan 030006)
[2](Department of Computer,Shanxi University,Taiyuan 030006)

Abstract: Collocations play an important role in parsing and verbs are the nucleus and precondition of parsing.This paper orients Chinese information processing,collocations.It concludes automatic matching-gotten rule by analyzing real text,uses statistical model,and studies character of matching type distribution and location attributed character of matching verbs.Then it successfully extracts the words and words matching in the material selected.Its precision rate arrives 75% and recall rate arrives 64%.

Keywords: collocation,corpus,Chinese information processing

1　引言

词语搭配是十分重要的语言知识。动词和动词性结构在语言研究中尤其重要。这是因为,动词与其他词类相比,内部最为复杂,动词在句法结构中的活动能力最强,大部分词类都要跟它发生一定的组合关系。在一般的句子里,动词常常占据着中心地位,它是句子的结构联系和语义联系的中心。

国外比较早开始搭配研究的是 Choueka 等[1],国内是孙茂松等[2]。该文的研究目标是:对一个经过分词和词性标注处理的汉语句子,通过自动分析确定句子中动词与动词是否是合法的的搭配,并确定其类型。例如,对于句子:

企业/n 改革/vn 继续/v 深化/v,/w

目标是:通过扫描句子,找出句子中的动词改革、继续、深化之间是否是合法搭配并确定搭配类型,此句子的处理结果是:改革继续(主谓搭配)、继续深化(动补搭配)。

由于动词与动词构成的搭配情况非常复杂,因而发现搭配的准确率较低。主要有以下难点:

(1)同样的动词与动词,在不同的语境中构成的搭配类型不同。

如:我们/r 的/u 行动/vn 促进/v 他们/r 的/u 竞争/vn 意识/vn(主谓搭配)

我们/r 以/p 实际/a 行动/vn 促进/v 改革/v 的/u 不断/d 深化/v。/w(非搭配)

(2)构成同样搭配类型的动词与动词,又无严格的规律可循。

如:保证/v 治污/vn 设备/n 正常/ad 运转/v,/w(动宾)

国内/s 旅游/vn 人数/n 及/c 收入/n 也/d 比/p 上年/t 有/v 大幅/b 增长/vn。/(动宾)

(3)句子中出现多个动词时,动词与动词搭配的交叉现象。

如:支持/v 和/c 声援/v [非洲/ns 人/n 国民/n 大会/n]nt 领导/v 的/u 南非/ns 人民/n 反对/v 种族/n 隔离/vn 的/u 正义/n 斗争/vn。/w

该文面向中文信息处理,通过对北大 200 万语料的统计分析归纳了搭配自动获取规则,结合统计模型,研究了动词-动词搭配中各关系类型的分布特征以及搭配词语的位置分布特征,在此基础上成功地抽取出所选语料中的动词—动词搭配。其中抽取正确率为 75%,召回率为 64%。

2　动词与动词搭配现象分析

2.1　搭配的定义

对于什么是搭配,语言学家由于理论背景和应用目的的不同,存在着不同的理解。

该文采取的定义是:搭配是由两个或以上的有一定的语法关系和语义联系的词组成的一种重复出现的词语组合。

2.2　动词-动词搭配现象考察

对动词-动词搭配关系类型及搭配词语的位置分布情况的考察结果如表 1。

基金项目:国家 863 高技术研究发展计划项目(编号:2001AA114031)资助
作者简介:白妙青,女,讲师,硕士研究生,研究方向为自然语言处理。郑家恒,女,教授,主要研究领域:自然语言处理。

表 1　动词–动词搭配关系类型及搭配词语的位置分布情况表

	动宾	动补	连谓	并列	主谓	状中	定中
R1	373	58	13	27	59	16	16
R2	63	8	7	22	36	13	27
R3	59	4	3	7	55	12	53
R4	52	2	2	2	10	8	7
R5	29	1	1	2	8	7	5
总个数	597	73	26	60	168	56	108
所占百分比	54.8%	6.70%	2.38%	5.51%	15.4%	5.14%	9.92%

其中 $R(X)$ 是指动词–动词搭配的窗口，窗口即为动词–动词搭配中两个动词之间词的个数。

从上表可以看出，动词宾语搭配词在 $R1$ 位置上出现次数明显较多，在其它各位置上的分布相对比较平均，只是 $R5$ 位置上的分布略有减少，反映出动宾搭配分布离散性强的特点。因此，可以认为动宾搭配窗口开在〔0，+5〕范围内是必要的，其它各类型窗口类推。最后得出各类型适宜窗口分别是：

动宾搭配〔0，+5〕，如：采取/v 实际/a 的/u 行动/vn

动补搭配〔0，+3〕，如：拿/v 出/v 诚意/n

连谓搭配〔0，+3〕，如：学习/v 掌握/v 先进/a 科学/n 文化/n 知识/n，/w

并列搭配〔0，+3〕，如：为/p 企业/n 的/u 改革/vn 和/c 发展/vn 建功立业/i。/w

主谓搭配〔0，+5〕，如：人民/n 生活/vn 进一步/d 改善/v。/w

状中搭配〔0，+5〕，如：江/nr 泽民/nr 同志/n 最近/t 强调/vd 指出/v，/w

定中搭配〔0，+3〕，如：要/v 大力/d 开展/v 劳动/vn 竞赛/vn、/w

3　动词–动词搭配训练策略

3.1　动词–动词搭配标注规则库

通过分析子语料，归纳了动词–动词搭配规则，由于篇幅所限，列出了其中几条规则，形式化描述如下：

令 W_l、W_r 分别表示搭配左项和搭配右项，设“W_L–W_r”序列中间包含 x 个词，依次为 W_1，W_2，……W_{x-1}，W_x，它们对应的词性分别为 T_1，T_2，……T_{x-1}，T_x。

(1) $R1$：x>=1 and W_1=‘的’

例如：制定/v 了/u 中国/ns 跨/v 世纪/n 发展/v 的/u 行动/vn 纲领/n

(2) $R2$：x=1 and T_1=‘p’

例如：18/m 家/q 因/p 治理/v 无望/v 被/p 责令/v 关停/v

(3) $R3$：x=2 and T_1=‘p’ and T_2=‘d’

例如：竟/d 能/v 以/p 几乎/d 令/v 人/n 难以置信/i 的/u 精确度/n 对/p 最/d 轻微/a 的/u 二氧化碳/i 含量/n 改变/v 作出/v 反应/v

(4) $R4$：x=1 and T_1=‘f’

例如：都/d 是/v 在/p 保盟/n 的/u 帮助/v 下/f 到达/v 抗日/v 根据地/n 延安/ns 的/u

(5) $R5$：x>=2 and T_1=‘f’ and T_2=‘的’

例如：环境/n 的/u 改变/v 引起/v 动物/n 需要/v 上/f 的/u 改变/v

3.2　统计模型

分析了一定数量的统计数据后，在此基础上构造了 V–V 搭配自动获取的统计计算模型，力图采用统计量，更全面考虑搭配的各项性质。选用的统计量是跨度搭配概率。

动词与动词跨度搭配概率：搭配通常是具有一定结构的，搭配词并不是完全等概率的分布在各个位置，而总是倾向于出现在某一个或某几个位置上。对某些搭配，所辖的两个词之间允许有间隔，甚至调序，但仍保持一定的结构关系（Smadja，1993）。动词–动词构成搭配在某几个位置上的出现概率远远高于其它位置。

任意位置上的概率定义为：

$$P(Sp)=\frac{Count(LOC=j)}{f(w)}$$

如果计算搭配动词在某个具体位置出现的概率，会导致严重的数据稀疏；因此采用各位置概率的均值，即：

$$P(Sp_{i=j}^{n})=\frac{\sum_{j}^{n}P(Sp)}{N} \quad (1)$$

其中，j、n 表示当前区间的起点和终点，N 表示区间内有效位置数量。

设阈值 w，当概率小于 w 时，V–V 不是合法的动词搭配，通过实验确定 w=0.2（见表 2）。

表 2　跨度搭配概率表

动词	动–动跨度	频率	概率
建立	1	21	0.875
建立	2	25	0.891
建立	3	18	0.816
建立	4	13	0.721
建立	5	3	0.106
建立	6	2	0.088

表 2 中可以分为两个区间：1~4 及 5~6。

3.3　动词与动词搭配类型自动标注算法

(1) 如果扫描到文件尾，那么转到(5)，否则处理当前扫描的句子：取出句中所有 V–V 及其中的各个词、词性及位置，存入预搭配库；

(2) 统计动词与动词间的跨度信息，利用公式(1)计算区间概率；

(3) 利用动词—动词搭配否定规则、V–V 跨度搭配概率，把预搭配库中不合法动词—动词搭配滤去，存入缩减库；

(4) 利用动词—动词搭配肯定规则及其位置分布特征判断搭配类型，存入搭配库；转到(1)；

(5) 结束。

例如：推动/v 两岸/n 经济/n 文化/n 交流/vn 和/c 人员/n 往来/vn，/w

把句子中的两个动词对推动……交流、交流……往来扫描到预搭配库（见表 3）。

表 3　预搭配库

词	词性	位置
推动	V	
两岸	N	1
经济	N	2
文化	N	3
交流	Vn	
交流	Vn	
和	C	1
人员	N	2
往来	Vn	

统计到的信息见表 4。

跨度搭配概率大于 0.2,经过(3)将两个搭配对存入缩减库,最后经过(4)将搭配结果存入搭配库(见表 5)。

表 4 信息库

动词	动词	跨度搭配概率
推动	交流	0.33
交流	往来	0.25

表 5 搭配库

搭配左词	搭配右词	搭配类型
推动	交流	动宾
交流	往来	并列

4 实验及分析

4.1 语料选取及对语料的预处理

这里所采取的语料来源于北大计算语言所的 200 万语料。该语料是经过了分词和词性标注的熟语料。规则及例句中用到的词性标记:a:一般形容词,c:连词,d:副词,L:习用语,j:名词性简称和略语,m:基数词,n:名词,f:方位词,ns:地名,p:介词,q:量词,r:代词,u:助词,vd:副动词,vn:名动词,v:动词。

以“小句”为单位,抽取出其中包含两个或两个以上动词的句子,形成子语料,而没有以“句子”为单位。“小句”是指由标点(逗号、分号、冒号、句号、问号、感叹号及省略号)将文本分割成的句法单位。这是因为:

(1)跨越小句范围的动词搭配为数很少。在对约 200 万词的语料进行手工标注的过程中,只发现了几十例。

(2)倘若以句子为单位,会增加很多不必要的竞争。

4.2 实验结果

这里用统计到的数据,对 200 万语料中的 60000 句进行了封闭测试,表 6 是自动识别的实验结果。

表 6

识别出的搭配总数	正确的搭配数	正确率
42000	31500	75%

$$\text{正确率}=\frac{\text{实际抽出的正确的动词-动词搭配总数}}{\text{实际抽出的动词-动词搭配总数}}\times 100\%$$

应识别的搭配总数	正确的搭配数	召回率
49500	31680	64%

$$\text{召回率}=\frac{\text{实际抽出的正确的动词-动词搭配总数}}{\text{应该抽出的动词-动词搭配总数}}\times 100\%$$

4.3 典型识别错误分析

(1)利用前面总结的抽取规则,虽然能将语料中大部分动词—动词搭配抽取出来,但由于规则库的规模小,还是有一些正确的动词—动词搭配被过滤掉。

例如:

致以/v 诚挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/vn !/w

其中,“问候”、“祝愿”为并列搭配,但是却被过滤掉。

(2)由于规则库的局限性,在抽出的动词—动词搭配中也有一些动词—动词搭配不合法。

例如:

我们/r 即将/d 以/p 丰收/vn 的/u 喜悦/an 送/v 走/v 牛年/t,/w

其中,“丰收……送”为不合法搭配,但是系统却并没有把它过滤掉。

5 结束

论文通过对真实文本的统计分析归纳了搭配自动获取规则,与统计相结合的方法将所选语料中的动词—动词搭配成功地抽取出来,抽取正确率为 75%,召回率为 64%。但是由于汉语言变化万千,还有一些规则没有考虑到,而且语料中对词性标记也存在一定的错误,因此在抽取动词-动词搭配的过程中,出现上述错误。需要进一步探索和研究的是,利用机器学习的方法不断扩大及改善规则库以及从语义层再进行研究,以提高该系统的召回率和正确率。(收稿日期:2004 年 2 月)

参考文献

1.Y Choueka,S T Klein,E Neuwitz.Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus[J].Journal of the Association for Literary and Linguistic Computing,1983;4(1):8~34

2.孙茂松,黄昌宁,方捷.汉语搭配定量分析初探[J].中国语文,1997;(1):29~38

3.车万翔.面向依存文法分析的搭配抽取方法研究[C].见:全国第六届计算语言学联合学术会议,2001:102~107

4.孙宏林.词语搭配在文本中的分布特征[C].见:黄昌宁编.1998 中文信息处理国际会议论文集,北京:清华大学出版社,1998:67~72

(上接 45 页)

(3)UMKDSS 刻画了结构化数据挖掘与非结构化数据挖掘之间的关系,指出了复杂类型数据挖掘的重点和难点在于它的特征处理,即如何提取特征并降低特征空间的维数,使其近似转化为结构数据,然后再进行数据挖掘,这对于非结构化数据挖掘有重要的指导意义,因为它指明了非结构化数据转化为结构化数据的路径。

(4)知识发现内在机理 KDTIM 的三个机制诱导出的模型 KDD/ KDD*、KDK/KDK*, 以及由 KDD* 和 KDK* 协同而形成的 KD(D&K)是 UMKDSS 中的核心,因此,知识发现内在机理 KDTIM 的三个机制是 UMKDSS 的核心和基础。同时可以看出知识发现内在机理 KDTIM 研究方向是沿着知识模板 *OK* 轴方向进行的, 反映的是知识的不同抽象程度及其各种变化规律,因此归纳和演绎方法是其主要的研究手段。

(5)非结构化数据挖掘研究的思路是:先通过特征空间降维、特征提取和变换,把非结构化数据变成结构化数据,然后按照结构化数据挖掘方法来处理。(收稿日期:2004 年 6 月)

参考文献

1.李德毅.发现状态空间理论[J].小型微型计算机系统,1994;15(11):1~6

2.邸凯昌.空间数据发掘与知识发现[M].武汉大学出版社,2001:20~23

3.杨炳儒,周颖.知识发现系统内在机理[J].北京科技大学学报,2002;24(2):345~349

4.杨炳儒,唐菁.基于复杂类型数据的发现特征子空间模型(DFSSM)的研究[J].中国工程科学,2003;5(1):56~61

5.唐菁,杨炳儒.基于 Web 的文本挖掘[J].计算机工程与应用,2002;38(21):198~201