词语搭配研究中的统计方法*

邓耀臣
(大连海事大学 外语系,辽宁 大连 116026)

摘要:重点介绍了词语搭配研究中常用的三种统计方法的理念和实现方法,特别是对词语搭配研究中常见的MI值和T值的计算方法作了详细的介绍,并对每一种方法的优、缺点加以比较。

关键词:词语搭配;MI值;T值;统计方法
中图分类号:H03 **文献标识码**:A **文章编号**:1671-7031(2003)04-0074-04

Statistics in collocation study

Deng Yaochen
(Foreign Languages Department, Dalian Maritime Univ., Dalian 116026, China)

Abstract: This paper mainly introduces the statistics commonly used for collocation study, especially, the measurements of MI score and T score. The strong points and weak points of each method are also compared in the paper.

Key words: collocation; MI score; T score; statistics

一、引言

词语搭配指词与词的结伴使用这种语言现象,是一种高度因循性的词语组合,是词语间的典型共现行为[1]。在外语学习中,掌握和使用典型的词语搭配是学习本族语者的无标记语言的重要内容之一。词语搭配的典型性由搭配的概率属性决定,因为任何搭配都是可能的,只不过一些比另一些更为恰当[2]。因此,抽取"更为恰当"的、典型的搭配词成为词语搭配研究的一个重要方面。

本文重点介绍了在基于语料库的词语搭配研究中,运用概率信息和统计手段自动抽取典型词语搭配的三种主要方法:1)统计搭配词与关键词的共现频数;2)统计测量共现词项间的MI值;3)统计测量共现词项间的T值。本研究采用的语料库为上海交通大学JDEST语料库的一个子库,词容为1 185 594词次。语料库研究工具为Wordsmith V3.0软件。

二、词语搭配研究中常用的统计方法及其比较

1.搭配词频数统计

要统计某一节点词(node word)的搭配词在语料库中的出现频数,首先要对节点词进行带有语境的检索(KWIC),然后提取节点词在一定跨距(span)内与之共现的所有词项,最后统计共现词的频数。只有在语料库中与节点词共现达到一定次数的词项才有可能成为节点词的习惯性搭配词。本研究以词形power为例说明其典型搭配词的抽取过程。在本研究采用的JDEST语料库中节点词power的观察频数为1 567。以下是运用Wordsmith软件从该库中提取的节点词power的检

* 收稿日期:2003-07-04
作者简介:邓耀臣(1967—),男,山东烟台人,讲师,硕士研究生。

索行片断,界定跨距为－4/＋4。

1. But the combination of power , complexity , and newness seem
2. the need for computational power was evidenced in the
3. modes allows the processing power of the system to
4. users with maximum computing power at minimum cost leads
5. the least amount of power to drive the display
6. display requires little processing power itself .
7. a greater combined processing power and can handle more
8. returns in terms of power for the extra financial
9. relative decrease in equivalent power as the number of a
10. For example , the equivalent power of a 4 processor system
11. processors , and the equivalent power of a 10 processor system ,
12. access to substantial computer power is now economically feasible
13. model has universal computing power .
14. maximum temperature . With power at 1 1/2 D per unit ,
15. actually serves as a power limiter . Since it is
16. input speed varies , however power limiting can still be

运用 Wordsmith 软件的统计搭配词功能,我们提取了节点词power 在－4/＋4 跨距内的共现词共 2 147 个。表 1 显示的是频数最高的 20 个共现词:

表 1　power 频数最高的共现词

rank	collocates	frequency	rank	collocates	frequency	rank	collocates	frequency
1	The	870	8	for	184	15	are	73
2	of	486	9	nuclear	133	16	plant	73
3	and	339	10	as	97	17	or	67
4	to	271	11	be	97	18	on	65
5	A	254	12	by	89	19	system	63
6	In	243	13	at	87	20	from	61
7	Is	211	14	with	75			

共现词频数表能使研究者很清楚地看出哪一些词与节点词经常在一起使用,并使研究者很容易确定一些明显的词语搭配。如从表 1 我们可看出,nuclear , plant 和system 三个词在语料库中与power 反复共现,构成意义明晰的搭配,都表达了科技英语中的一些重要概念。

但是,通过共现词的频数确定搭配词的方法存在严重缺点。从以上检索行可看出,由于界定跨距忽略句子界限,power 的一些共现词与节点词没有语法限制关系,对节点词也没有任何预见作用。如第 15 行中的power . . . Since 和第 16 行中的However ,power 等。这些共现词落入跨距内完全是由语言使用的某种偶然因素造成的。在一般的语料库研究活动中,它们被称为偶然搭配词。这些偶然搭配词不是真正意义上的词语搭配,应当排除。另外,仅根据共现频数的高低,研究者还无法确定每一个共现词是否为显著搭配词。如表 1 中the 位于频数之首,是因为它与power 的相互预见、相互吸引力最强还是因为它能与所有名词连用而造成共现频数最高? 我们最关心的是,在特定的语境内,节点词power 对哪一些词产生了显著影响,以至于吸引它们与之构成典型搭配。如果都是显著搭配词的话,他们的显著性有什么不同? 要回答这些问题,我们就必须运用统计测量的方法,检验每一个共现词与节点词之间的相互预见和相互吸引程度,判断它们的共现在多大程度上体现了词语组合的典型性。

2. MI 值的统计测量

共现词显著性的测量方法通常有两种:MI 值和T 值。这两种方法都是通过比较共现词的观察频数(observed frequency) 和期望频数(expected frequency) 的差异来确定搭配序列在语料库中出现概率的显著程度[3] 。

期望频数是词语搭配研究中的一个重要概念。这一概念的提出基于这样一个假设:如果节点词对共现词没有吸引、预见影响的话,那么共现词在节点词特定跨距内出现的概率应该和在整个语料库中随机分布的概率一样。假设 x 和 y 分别

为语料库中随机分布的两个词,语料库的总词容为N,它们在语料库中出现的实际观察频数分别为$f(x)$和$f(y)$,出现概率为$P(x)=f(x)/N$和$P(y)=f(y)/N$,那么,如果搭配词y不受节点词x的吸引而与之共现(即:在x的特定跨距内出现)的期望频数应为:$f(o)=f(y)/N*[f(x)*2S]$(S为跨距)。

MI值(Mutual Information Score,互信息值)表示的是互相共现的两个词中,一个词对另一个词的影响程度或者说一个词在语料库中出现的频数所能提供的关于另一个词出现的概率信息。MI值越大,说明节点词对其词汇环境影响越大,对其共现词吸引力越强。因此,MI值表示的是词语间的搭配强度。MI值的计算公式为:$I(x,y)=\log_2[P(O)/P(E)]=\log_2[f(x,y)*N]/(f(x)f(y)*2S)]$。如果$x$和$y$之间存在真正的连结关系,那么观察概率将远大于期望概率,结果为$I(x,y)>0$。如果两个词相关程度不高,那么观察概率接近期望概率,结果为$I(x,y)\approx 0$。如果$I(x,y)<0$,说明其中一个词出现时,另一个词不出现,即二者呈互补分布[4]。例如nuclear一词在语料库中的观察频数$f(y)=493$,与power共现的频数$f(x,y)=133$,那么其MI值为:$I(x,y)=\log_2[f(x,y)*N]/(f(x)f(y)*2S)]=\log_2[(133*1\,185\,594)/1\,567*193*8]=10.67$。基于语料库的词语搭配研究中通常把MI值等于或大于3的词作为显著搭配词[3],所以nuclear和power能构成显著搭配。表2是节点词power在JDEST语料库中MI值最高的20个搭配词:

表2 power在JDEST语料库中MI值最高的20个搭配词

rank	collocates	fy	fxy	Mscore	rank	collocates	fy	fxy	Mscore	rank	collocates	fy	fxy	Mscore
1	Horse	13	10	12.18	8	Plants	207	55	10.65	15	Capacitors	20	4	10.24
2	Throttles	6	4	11.98	9	Dissipation	27	7	10.62	16	Tie	20	4	10.24
3	Excursion	11	5	11.43	10	Stations	74	19	10.6	17	Tremendous	20	4	10.24
4	Watts	18	6	10.98	11	Actuated	20	5	10.56	18	Supplies	57	11	10.19
5	Nuclear	493	133	10.67	12	Chassis	23	5	10.36	19	Reactive	53	10	10.16
6	Fueled	15	4	10.66	13	Versions	24	5	10.3	20	Bundle	43	8	10.14
7	Outage	15	4	10.66	14	Stroke	59	12	10.27					

MI值主要是通过测量共现词的非随机性(non randomness)来体现词语搭配的显著性。上表中horse和power共现的MI值是12.18,这意味着horse和power的实际共现频率比偶然概率高出$2^{12.18}=4\,640.29$倍。因此,我们可以断定二者之间有较强的连结关系,能构成典型搭配。

MI值可清楚表现共现词间的相互吸引程度,它可以帮助我们确定把哪些词作为节点词的可能搭配词而重点加以研究。但是MI值高的搭配词不一定和节点词共现的频数就高。以表2中的throttles一词为例。尽管throttles在JDEST语料库中频数较低(仅出现6次),但从MI值(11.98)来看二者搭配显著,这是因为throttles在语料库中几乎都是与power结伴共现(4次)。这说明MI值表示的词语连结信息并不总是可靠。如果一个词在语料库中的出现频率较低,而出现时又多与节点词共现,那么二者的MI值肯定很高。这也说明对于语料库中的低频词(频率小于10),MI值信度较低,因为我们不能确定这一结果是源于二者的真正关联还是源于语料库的特殊本质。因此,在词语搭配抽取的研究中,除了计算搭配强度外,还有必要对共现词的显著性进行假设检验,以获得有关典型搭配的更多证据。常用的检验方法为t检验。

3.T值的统计测量

T值是根据假设检验中的t检验计算得来的。假设检验主要通过检验某一样本的平均数与正态分布总体的平均数之间的差异是否显著来断定该样本取自总体的可能性有多大,或者说二者之间的差异是否由偶然性造成。在词语搭配研究中,我们要检验的就是在由节点词构成的小文本中两个词的共现频数与期望频数是否存在显著性差异。运用t检验断定搭配词的显著性时,首先形成零假设:两个共现词之间没有联系,不能构成搭配,然后以标准差来衡量观察频数和期望频数的差异是否达到显著性水平。

计算T值时首先要计算搭配词在小文本中分布的标准差。计算公式如下:

$$SD = \sqrt{f_{(y)}/N \times (1 - f_{(y)}/N) \times f_{(x)} \times 2S}$$

T 值计算公式为

$$t = (f_{(o)} - f_{(e)}) / SD$$

如果T 值小于显著性水平为 0.05 的关键值(critical value) 1.65,我们就保留零假设。否则,可以推翻零假设,而得出两者可以构成显著搭配的结论。通常情况下,我们把T 值等于或大于 2 的搭配词作为显著搭配词。我们仍以nuclear 为例,计算其T 值。

$$f_{(e)} = f_{(y)}/N \times f_{(x)} \times 2S = 493/1\,185\,594 \times 1\,567 \times 8 = 5.21$$

$$t = [f_{(o)} - f_{(e)}] / SD = [133 - 5.21] / \sqrt{5.21 \times (1 - 0.000\,4)} = 55.98$$

由于T = 55.98 > 1.65,所以我们有 95% 的把握推翻零假设,而得出结论:nuclear 和power 能构成显著搭配。表 3 是节点词power 在JDEST 语料库中T 值最高的 20 个搭配词:

表 3　power 在JDEST 语料库中T 值最高的 20 个搭配词

rank	collocates	fy	fxy	tscore	rank	Collocates	fy	fxy	tscore	rank	Collocates	fy	fxy	tscore
1	nuclear	493	133	55.98	8	reactor	664	49	15.85	15	Reactive	53	10	12.61
2	plants	207	55	35.7	9	stroke	59	12	14.4	16	Dissipation	27	7	12.57
3	plant	466	73	30.67	10	excursion	11	5	14.32	17	Generation	180	19	12.39
4	horse	13	10	26.6	11	Output	503	38	14.17	18	Consumption	66	11	12.33
5	supply	290	44	23.38	12	Supplies	57	11	13.39	19	Unit	463	31	11.8
6	stations	74	19	20.6	13	Watts	18	6	13.32	20	Full	360	26	11.38
7	station	122	22	18.24	14	Solar	222	22	12.83					

观察表 3 可看出该表中的搭配词既不与表 1 频率表中的共现词完全相同,也与表 2 MI 值表中的搭配词有所区别。如定冠词the 与power 共现的MI 值(5.91) 和T 值(-1.78) 都很低,这说明the 位于频数之首在很大程度上是因为它是一个高频词,而不是因为与power 有较强的搭配力。MI 值表中的throttles 一词,尽管搭配力较强,但与节点词共现频数太少,缺少足够证据,因此在t 检验中也被过滤掉。而对于nuclear , plants ,horse , supply , stations ,excursion , watts 等词,我们可以准确地断定它们是power 的典型搭配词,因为它们的MI 值和T 值都达到显著性水平。

三、结语

本文重点介绍了词语搭配研究中常用的三种统计方法的理念和实现方法,并对它们在词语搭配研究中的不同功能进行了区别。分析表明共现频率可以使研究者很容易找到一些明显的搭配词(“nuclear power” , “power plant” 等) ,MI 值测量的是搭配强度,它有助于识别科技术语和固定词组(“outage power” , “access power” 等) ,而T 值反映的是对显著搭配词的把握性(certainty) ,能使研究者有把握地确定与节点词共现频数较高的显著搭配词(“horse power” , “supply power” , “power station” 等) 。在实际的词语搭配研究活动中,我们可以将MI 值和T 值结合使用,如果一个搭配词的两种统计量都达到显著性水平的话,那么它肯定就是节点词的显著搭配词。

参考文献:

[1] Firth J R. Papers in Linguistics 1934—1951[M]. London: Oxford University Press, 1957.

[2] Sinclair J. Beginning the study of lexis[A]. In: Bazel C E, Catford J C, Halliday M A K, et al. In memory of J. R. Firth[C]. London: Longman, 1966.410-430.

[3] Hunston S. Corpora in Applied Linguistics[M]. Cambridge: Cambridge University Press, 2002.

[4] Church K, et al. Using statistics in lexical analysis[A]. In: Zernik U. Lexical Acquisition: Exploring On-line Resources to Build a Lexicon[C]. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.

[5] Church K, Hanks P. Word association norms, mutual information and lexicography[J]. Computational Linguistics, 1990(16): 22-29.