DOI:10.13852/j.cnki.jshnu.2006.06.016

中图分类号:H0-05 文献标识码:A 文章编号:1004-8634(2006)06-0117-(06)

词项搭配的定量分析方法

汪腊萍

(上海师范大学 外国语学院,上海 200234)

摘 要: 文章阐明了语言学家如何将信息理论中的互信息熵(MI值)和统计学及概率理论中的零假设T检验值(T值)实际应用于测量词项共现中两词项间的搭配强弱(Collocability)和判断搭配强弱的置信度(Confidence Level)。并通过例词分析,说明了利用MI值和T值可以从大量的词项共现中分辨出高频典型搭配,阐明了MI值和T值所提供量化标准的客观性和科学性,并指出藉此得到的有关高频典型搭配资料,对于提高外语学习者的搭配意识和搭配知识、优化外语教学效率和效果,将是不可或缺的资源。词项搭配研究对于语义识别、词典编撰、教材编写、ESP大纲设计、比较语言学研究等也有较大的指导意义。

关键词: 统计学分析方法;词项搭配;互信息熵(MI值);T检验值;关联程度;置信度

1. 问题的提出

近几年来词项搭配研究一直是国内外语料库语言学家最感兴趣的研究领域之一。所取得的研究成果可从最近出版的搭配词典、搭配教材和搭配研究文献及搭配信息网站中窥见一斑。我国语料库研究人员对词项搭配的定性和定量研究(冯和汪1999a和b;杨惠中2002;)使得语言教育者和学习者从词项搭配的庞杂中有了一些头绪,通过对可信赖语料库所提供的统计资料的分析,对互信息熵值(简称MI值)搭配词表和T检验值(T值)搭配词表的互为补充的对照研究,可以获得所选中心词所有典型搭配词的语义特征,得到对该词项及其所有搭配词的全息图片信息。

虽然MI值和T值所提供的搭配词表已在诸多研究人员的研究文献中得到应用(冯和汪1999a和b;杨惠中2002:115-6,201-2,205,207),而且应用于语料库语言学研究的MI值的信息学理论和统计学原理也得到了较为清晰的阐述(冯和汪1999a),但是,至目前为止,我国尚无人将概率理论中零假设T检验值的语料库语言学应用价值从统计学原理的角度加以阐明。本文结合语料库语言学、信息论和统计学原理,对真实语言资料中的词项共现进行数理统计分析,旨在进一步阐明用MI值进行词项搭配定量研究时所获得的搭配资料及所面临的问题,着重介绍并分析T值的统计学理论和应用价值,及其对MI值的修正和完善。旨在分析说明语料库语言学家为什么可以将统计学的MI值和T检验理论应用到词项搭配的定量研究中。

收稿日期:2006-09-06

作者简介:汪腊萍(1965-),女,湖北天门人,上海师范大学外国语学院副教授,主要从事语料库语言学及计算语言学研究。

2. MI值在词项搭配定量分析中的应用

2.1 理论基础及实际应用

互信息熵(Mutual Information)(MI值)(Fano1961)是建立在信息理论基础上的一个概念,现在用它来测量中心词(node word)和其搭配词(collocate)之间的相互关联程度(association strength)或搭配强弱(collocability)。英国伯明翰大学COBUILD语言研究中心采用了一种易为大众理解和接受的MI值计算方法。其计算公式为:

$$MI(x,y)=\frac{f_{obs.}(x,y)}{f_{exp.}(x,y)}$$ (*Church* 等1990;冯和汪1999*a*)

这里,中心词为x,定长内(中心词前后各4个词)搭配词为y;中心词与搭配词的观测共现频数为$f_{obs.}(x,y)$,零假设下中心词与搭配词的期望共现频数为$f_{exp.}(x,y)$。

MI值所反映的是所观测到的共现频数与所期望的共现频数在多大程度上不一致。即:中心词x的出现给搭配词y的出现带来了多大信息。一般而言,MI值越大,说明中心词x对搭配词y的出现的影响就越大(冯和汪1999a & b)。

2.2 词例分析:*save* 的搭配词分析

根据以上公式和步骤,可以算出*save*定长内搭配词的*MI*值(见表1)。表1资料由五栏构成:(1)定长内的搭配词(y);(2)每一个搭配词在语料库中的总频数($f(y)$);(3)中心词与搭配词的观测共现频数$f_{obs.}(x,y)$;(4)零假设下中心词与搭配词的期望共现频数$f_{exp.}(x,y)$;(4)降序而排的所有搭配词的MI值$MI(x,y)$。

表1　按MI值降序排列的*save*搭配词词表

$f(x)=f(save)=611$　　$N=5771842$

(y)	$f(y)$	$f_{obs.}(x,y)$	$f_{exp.}(x,y)$	$MI(x,y)$
fund	318	8	0.2701	29.6296
lives	507	11	0.4311	25.5814
seed	278	6	0.2409	25.0000
images	676	13	0.5732	22.8070
money	1592	25	1.3525	18.5185
print	736	11	0.6217	17.7419
fuel	459	5	0.3942	12.6839
document	527	6	0.4513	13.2949
load	647	7	0.5478	12.7783
file	2592	28	2.2009	12.7221
space	1305	14	1.1107	12.6047
campaign	489	5	0.4095	12.2100
files	2260	17	1.9070	8.9145
image	1253	8	1.0612	7.5386
energy	1510	9	1.2851	7.0033
disk	2718	14	2.3027	6.0798
screen	2245	11	1.9032	5.7797
lot	1458	7	1.2457	5.6193
copy	1555	6	1.3239	4.5321
time	8877	31	7.5243	4.1200
can	24633	79	20.8607	3.7870
you	43177	131	36.5736	3.5818
per	1711	5	1.4490	3.4507
children	4484	12	3.7895	3.1666
data	2765	7	2.3309	3.0003

2.2.1 *save* 搭配词的语义特征

表1所列的是在 *Tim John* 语料库中、*save* 搭配词中 MI 值大于 3 的所有搭配词(冯和汪 1999*a*)。这些搭配词大部分都是名词,很可能是作 *save* 的宾语。由表1可知道这些搭配名词主要呈现出两类语义特征。一类是人类极其珍贵的东西,比如 *life*、*money* 以及 *time*。另一类是与计算机领域有着或多或少联系的词群,比如 *image*、*print* 以及 *data* 等。从理论上讲,英文中有很多词汇可用来表达将某东西保存起来,以备后用,比如 *keep*、*preserve*、*remain*、*store*,但没有一个词含有 *save* 所蕴涵的联想意义。根据 *Longman Dictionary of Contemporary English* (*Summers* 1995), *save* 所列的第一个词义是 *make someone or something safe from danger, harm, or destruction*。所以,当人们选择 *save* 这个词去拯救、节省或保存某人或某物时,他们有意无意地昭示了所接宾语对他们的重要意义。

基于 MI 值基础上的搭配词研究揭示了 *save* 及其搭配词的语义场特征。也就是说,学习者在使用 *save* 时应记住其搭配词多与计算机领域及 *life*、*time*、*energy* 等人类极其珍惜的对象有关系。因此,任何一个新诞生的搭配词只要有这样的语义特征,便是合适的、可接受的。

2.2.2 *save* 搭配词的语法特征

在 *save* 的搭配词项表中,有两个语法词项非常醒目:人称代词 *you* 和情态动词 *can*。对这两语法搭配词项的进一步研究表明,在所有 *save* 和 *can* 的搭配中有 72.6% 的搭配定长中含有人称代词 *you*。

从数据统计分析来看,*you can save* 或 *can save you* 这一三词搭配现象十分突出。这种突出的语法词项搭配可能来自于目前倡导的英语交际功能,尤其是商务英语交际活动中所一贯推崇的"You - Attitude"(O. Locker 1998:34)。即无论是在书面或口头讨论任何问题时,读者或听众或顾客——*You*,是作家或商家所关注的第一中心。而 *save* 这个词所蕴涵的联想意义,与读者第一或顾客至上的观念互相呼应,便有了 *you can save* 和 *can save you* 这一实义词与两语法词项的典型搭配。

2.2.3*save* 的专业术语搭配

表1还展示了一个典型技术术语搭配:*save* / *seed*。搭配 *save* / *seed* 拥有较高的 MI 值(MI = 25),主要是因为搭配词 *seed* 的总频数相对是最低的,其共现频数也相对较低,在这种情形下,我们说 *save/seed* 不能被看作是常用的、典型的词项搭配,但它可被看作是有较强的专业领域性的词项搭配(冯和汪 1999a 和 b)。

2.3 MI 值存在的问题及解决方法

用 MI 值识别出来的搭配词种类繁多,同时又呈现出一定的语义特征。这正是 MI 值的统计意义:测量词项间搭配程度的强弱(collocability)。但是,当中心词与搭配词的观测共现频数 $f_{obs}(x, y)$ 较小或搭配词在语料库中的总频数 $f(y)$ 相对较小时,MI 值是不太可靠的。这就是为什么 *save* / *seed* 不能被认为是典型的词项搭配,只能被看作是一种较强的专业领域词项搭配。

因此,对于 $f_{obs}(x,y)$ 或 $f(y)$ 较小的搭配词来讲,即使 MI 值较大,也不能冒然断言它与中心词的搭配性较强。这时,T 值可以帮助我们消除 MI 值遗留的困惑,剔除那些可能观测到的但不典型的词项组合,突出真正常用的、有代表性的词项搭配。

3. T 值在词项搭配定量分析中的应用

3.1 理论基础

T 值是用来检验两个母体有无差异的二阶统计值。它反映的是两个母体的相对差异。若要比较不同系统下两母体间的差异,则要比较它们的 T 值,即用均方差来衡量个体与均值的差异(王康 1988:273 - 5)。

将 T 检验理论用于词项搭配的定量研究是无数语言学家和语言研究人员长期艰辛努力的结果。在这些先驱者中有 Church 和他在美国电话电报公司研究中心的研究人员。他们的研究为将 T 检验理论用于搭配研究的实际应用起到了搭桥铺路的作用。

3.2 实际应用

根据 T 检验统计理论的原理,假设两词项,x 和 y,在某语料库中共现概率为 $P(x, y)$,各自单独出现的概率为 $P(x)$ 和 $P(y)$,那么所观测到的共现概率 $P(x, y)$ 与随机共现的偶然概率 $P(x)$ $P(y)$ 之间的 T 值为:

$$T(x,y)\approx\frac{P(x,y)-P(x)P(y)}{\sqrt{\frac{P(x,y)}{N}}}$$ (*Church* 等 1991)

这里,共现概率 $P(x,y)$ 由 x 和 y 两词项在指定的语料库中所观测到的共现频数 $f(x,y)$,除以该语料库的大小 N 而得。$P(x)$ 和 $P(y)$ 由语料库中 x 和 y 单独出现的频数 $f(x)$、$f(y)$ 除以该语料库的大小 N 而得。$P(x,y)-P(x)P(y)$ 反映的是所观测到的共现概率与随机共现的偶然概率或所期望出现的共现概率(Sinclair 1991:70)之间的绝对差异,它受 $P(x,y)$、$P(x)$、$P(y)$ 数值本身大小的影响。而值 $\frac{P(x,y)-P(x)P(y)}{\sqrt{\frac{(P(x,y))}{N}}}$ 所反映的是观测共现概率与偶然共现概率用均方差来衡量的相对差异。

根据T检验的统计意义,两词项 x、y 的T值反映的是两词项间搭配强弱的相对差异。从统计学的角度来看,1.65个均方差的差别是判断两词项搭配是否有意义的最低临界值。1.65个均方差表明我们有95%的把握断言观测共现概率 $P(x,y)$ 与偶然共现概率 $P(x)\ P(y)$ 的差别是客观存在而非偶然巧合。也就是说,对两词项 x、y 的共现,我们有95%的把握说它们是有意义的搭配。不同的T值给我们不同程度的信心来断言两个或更多的词项组合是否有意义的搭配。比如:

$T(x,y)\geq 3.09$,表明词项 x、y 的共现有0.1%的可能是偶然相遇;换句话说,我们有99.9%的信心宣称词项 x、y 的共现关系是有意义的;

$T(x,y)$ 3.30 表明词项 x、y 的共现有0.05%的可能是偶然相遇;换句话说,我们有99.95%的信心宣称词项 x、y 的共现关系是有意义的。(李沛良 1987:57-61)。

假设判断词项组合是否有意义的词项搭配的临界T值定为3.09个标准差。那么,如果 $T(x,y)<3.09$,则表明 $P(x,y)\approx P(x)P(y)$,即 $P(x,y)$ 和 $P(x)P(y)$ 的差别并不明显到足以断言 x、y 之间存在较强的联系。换句话说,x、y 两词项共现更有可能是自由组合,而不太可能是典型搭配。

如果 $T(x,y)\geq 3.09$,则表明 $P(x,y)$ 和 $P(x)P(y)$ 之间的差别是如此之大以至于我们有99.9%的把握宣称 x、y 之间存在较强的联系。它们很有可能是典型的词项搭配。

3.3 *词例分析:save 的搭配词分析*

表2所列的是在同一语料库(Tim John 语料库)中、以降序T值排列的save搭配词词表。该表只包括了T值大于3.30的搭配词。现在我们可以比较一下用MI值挑选出来的搭配词和用T值挑选出来的搭配词到底有什么区别。

3.3.1 save搭配词的语义和语法特征

所有T值挑选出来的搭配词也可分为两类:一类为实义搭配词;另一类为语法搭配词。实义搭配词同由MI值挑选出来的搭配词有一定的覆盖性,比如都有 *data*、*time*、*file*、*money* 之类的词项。这说明这些词项同 *save* 确实存在很强的搭配关系,它们共同构成了 *save* 的典型常用搭配。

3.3.2 MI值搭配词表和T值搭配词表对照分析

排在T值前十位的搭配词中语法搭配词就占了8个。一般而言,T值挑选出来的搭配词本身都是出现频度非常高的词汇,比如像介词、人称代词、冠词、小品词等语法词项。因为T值相对来讲对原始频数比较敏感,因此我们在利用T值分析语法词项与中心词的搭配是否有意义时一定要特别小心谨慎。在本词例分析中,T值最高的语法搭配词项是 *to*。我们可能会冲动地认为:动词 *save* 可能会总在不定式 *to* 的前面或后面。但是,如果我们再看看 *save* 的MI值搭配词表,就会发现:词项 *to* 根本不在MI值搭配词词表之列!事实上,MI(*save*, *to*)值只有1.8621!它表明这两词项的语义关联程度很弱。因为语法搭配词项自身都是高频词,它们有很强的与其它词项共现的趋势。虽然相对而言有较高的 T 值,但并不说明它们同中心词的关联程度很强,比如 *save a*、*save the* 等。只有MI值和T值相互对照,才能去伪存真,识别出真正有意义的词项搭配。比如 *save* 同 *you*、*can* 的MI值和T值都很高,那么,我们可以十分肯定地说,*save* 对语法词项 *can*、*you* 确实心存偏爱。

表2 按 T 值降序排列的 *save* 搭配词词表

$f(x)=f(save)=611 \quad N=5771842$

(y)	$f(y)$	$f(x,y)$	$T(x,y)$
to	171243	270	15.3270
the	344151	233	12.8475
you	43177	131	11.0457
and	214349	141	9.9609
can	24633	79	8.5948
a	166334	87	7.4371
it	52038	61	7.1040
data	2765	7	6.7069
in	162586	59	5.4375
time	8877	31	5.3988
file	2592	28	5.2396
money	1592	25	4.9662
from	26875	30	4.9571
that	49873	33	4.8243
or	34648	30	4.8067
will	14539	25	4.6918
for	40132	27	4.3775
as	40920	27	4.3614
on	45044	26	4.1626
would	10673	19	4.0994
files	2260	17	4.0650
your	42341	24	3.9824
of	147446	39	3.7423
space	1305	14	3.7047
by	28311	19	3.6704
disk	2718	14	3.6647
one	18864	17	3.6381
images	676	13	3.5857
children	4484	12	3.3269
lives	507	11	3.3114

3.4 两种定量分析方法的比较

MI 值,作为信息理论最基本的统计值,用于词项搭配定量研究时,它所测量的是一个词的出现在多大程度上预测着另一个词的出现。如果中心词和搭配词的共现频数 $f(x,y)$ 足够大,那么,MI 值的大小能够较准确地反映出该词项组合的关联强度和可搭配度(冯和汪 1999b);如果共现频数 $f(x,y)$ 或搭配词出现的频数 $f(y)$ 较小,如 *teachers / Steiner*(冯和汪 1999a), *save / seed* 等,即使 MI 值较大,也不能冒然断言它们的关联程度较强。因为从统计学的角度来讲,出现频率很小的事件,小样本观察时是不会发生的;即使观察到了,也不能轻易否定该事件发生机率很小这一事实(冯和汪 1999a;王康 1988)。

T 值的特点在于它更明显地突出了高频共现的搭配词,因而排除了技术术语、特殊癖好、偶然巧遇或化石化结构等词项组合。它可用来测量在判断词项组合可搭配度时的置信度(confidence level)(冯和汪 1999b)。但是 T 值挑选出的语法搭配词,如 *save the*, *save a* 等,即使有较高的 T 值,却并没有任何搭配意义。因为 *save a*, *save the* 的 MI 值往往较小,因而这种组合的搭配意义都很弱,而不太可能被认为是常用、典型的词项搭配。事实上,中心词与大多数语法词项的 MI 值一般都非常低,其原因是这些语法词项在英语中均属最常用词汇(Carter 1988),它们几乎有可能与任何一个相关的词项频繁共现,故这种共现组合一般不大可能具有搭配意义。

MI 值和 T 值存在着较强的互补关系。在进行词项搭配的定量分析时,若能结合 MI 值和 T 值搭配词词表对比分析,就一定能识别出可靠的、典型的、真正代表该词项搭配用法的词项搭配。一般说来,若某词项组合的 MI 值和 T 值都较大($MI \geq 3$; $T \geq 2.33$)(Church & Hanks 1990; Church et al 1991; 李 1987;冯和汪 1999b),则该词项组合可被认为是典型且常用的词项搭配。

总之,MI 值词表和 T 值词表可从语料库中将与任一中心词频繁共现的搭配词检索出来。经过适当地分析和比较后,从形形色色的各种词项共现中全方位地精选出该中心词的常用且典型的搭配词。使我们对中心词的认识不再是孤立的、片面的、干巴巴的,而是立体的、全息的、有血有肉的系统知识(冯和汪 1999b)。

5. 结束语

随着计算语言学、语料库语言学及其相关交叉学科的发展,现代语言学家对搭配现象的研究方法和研究成果都发展到了一种全新的领域。本文主要阐明 MI 值和 T 值用于词项搭配研究的统计学原理,旨在帮助对 MI 值和 T 值感兴趣的同仁在分析和利用 MI 值和 T 值搭配词表时建立更强的信心,更清楚地意识到这种量化资料的客观性和实用性,从而在各自的教学和研究领域中自

觉地使用和参考这些资料,进而改善我国英语词项搭配教学的效果和效率。

MI值和T值所提供的资料尤其在专门用途英语(ESP)研究领域、同义词使用的鉴别方面、两种或多种语言语义比较研究方面都有极大的潜力。这为广大的外语教师和研究人员在进行教材编写、语言研究等方面提供了极具价值的量化资料。

有一点需要特别指出的是:本项研究所用的语料库来自伯明翰大学Tim John教授及其同仁建立的500万词次的语料库(文中简称Tim John语料库)。本文的一切统计资料和研究结果都基于这一语料库产生的。

参考文献:

[1] *Carter, R.* 1988. *Vocabulary and Language Teaching*[M]. London: Longman Group UK Limited.

[2] Church, K. W., and Hanks, P. 1990. *Word association norms, mutual information, and lexicography*[J]. Computational Linguistics, 16(2), 22-29.

[3] Church, K.; Gale, W.; Hanks, P.; and Hindle, D. 1991. *Using statistics in lexical analysis*[A]. In Lexical Acquisition: Exploiting On-line.

[4] *Resources to Build a lexicon*[C], edited by Uri Zernik, 115-164. Lawrence Erlbaum Associates.

[5] O. Locker, K 1998. (*4th edition*) *Business and Administrative Communication*[M]. Beijing: Mechanical Industry Press.

[6] Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*[M]. Oxford: Oxford University of Birmingham.

[7] Summers, D. 1995. (3rd edition). *Longman Dictionary of Contemporary English*[M]. London: Longman Group Ltd.

[8] 冯跃进,汪腊萍.英语搭配行为的定量分析[J].国外外语教学,1999,(2).

[9] 冯跃进,汪腊萍.科比德在线演示版及其应用[J].外语学刊,1999,(4).

[10] 李沛良.社会科学统计研究[M].武汉:湖北人民出版社,1987.

[11] 王康.社会学词典[M].济南:山东人民出版社,1988.

[12] 杨惠中.语料库语言学导论[M].上海:上海外语教育出版社,2002.

Quantitative Approaches to Collocation Analysis

WANG Laping

(Foreign Languages College, Shanghai Normal University, Shanghai, 200234, China)

Abstract: On the basis of probability theory and information theory, this paper illustrates how two measures of significance, namely MI-score and T-score, are employed by linguists to measure the collocability and the confidence level of the association between word combinations. By means of case study, the paper demonstrates the objectivity of qualitative and quantitative evaluations of the approaches and of the results produced by these approaches in two corpora. The paper also points out the potential applications assisted by the two measures in language research areas, so as to cultivate language learners' collocation awareness, to accumulate their collocation knowledge and to improve the efficiency and effectiveness of collocation teaching and learning. The data obtained by these measures are also of guidance in the research areas like semantic categorization, textbook edition, ESP syllabus design, language production and comparative language studies. With an optimistic view to corpus-based computational linguistics, this paper indicates a more rigorous and revealing future for language learning and linguistic research.

Key words: statistical approach, collocations, MI-score, T-score, association strength, confidence level

(责任编辑:申 浩)