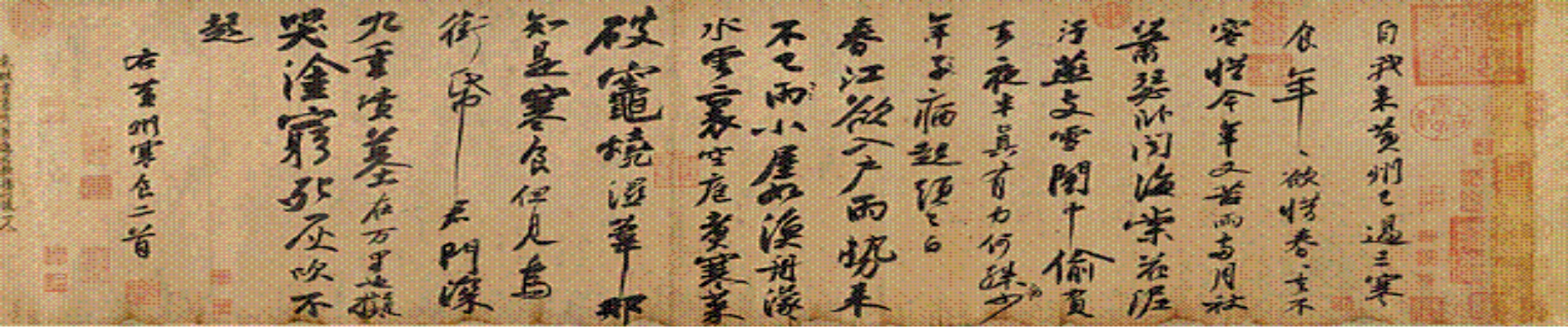


计算语言学
Computational Linguistics

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：**林衍凯**

Email:linyankai423@qq.com

郑重声明

- 此课件仅供选修清华大学计算机系研究生课《计算语言学》(70240052)的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中。未经孙茂松本人同意，任何人不得以任何方式扩散之（包括不得放到任何服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

第五章
搭配

5.1 Collocations

- Examples

  noun phrase:

  strong tea; weapons of mass destruction

  Phrasal verbs:

  make up

  Idioms:

  kick the bucket

- Habitual word usage

  a stiff breeze *? a stiff wind*

  a strong breeze *a strong wind*

  ? powerful tea *powerful drugs*

Google

"strong tea" Google 搜索

获得约 130,000 条结果 （用时 0.20 秒） 高级搜索

所有结果
图片
更多

网页
所有中文网页
简体中文网页

时间不限
2 月内

普通视图
图文并茂

更多搜索工具

小提示： 只搜索中文(简体)结果，可在 设置 指定搜索语言

high tea、hot tea、ice tea、strong tea的中文_百度知道

2008年7月7日 ... 5下的暑假作业，知道的帮帮忙，谢谢！ ... 1 下午餐, 茶点2 热茶3 冰茶4 浓茶 ... 傍晚茶点High tea盛行于普通大众。因为晚餐最早要在8点才能开始，作为一天 ...

zhidao.baidu.com/question/59529899 - 网页快照

Tea Obsession: When to drink your teas light and strong? - [翻译此页]

7 May 2008 ... Strong tea can clear excessive body heat, detox, nourish lung, ... Light tea is more beneficial than strong tea in terms of health ...

tea-obsession.blogspot.com/.../when-to-drink-your-teas-light-and.html - 网页快照 - 类似结果

Let it brew - strong tea may cut cancer risk - Times Online - [翻译此页]

27 May 2007 ... Let it brew - strong tea may cut cancer risk. Ruairi O'Kane. GRANDMOTHER did know best. Scientists have established that tea left to brew in

www.timesonline.co.uk/tol/life_and_style/.../article1845275.ece - 类似结果

Strong Tea - [翻译此页]

Strong Tea. Spirit of 1773 for Today ..... Copyright © 2010 Strong Tea - All Rights Reserved Powered by WordPress & the Atahualpa Theme by BytesForAll. ...

www.strongtea.org/ - 网页快照

Google

"powerful tea" Google 搜索

获得约 31,400 条结果 （用时 0.05 秒） 高级搜索

所有结果
更多

网页
所有中文网页
简体中文网页

更多搜索工具

Melaleuca - Powerful Tea Tree Oil Treatment for Acne, Staph and ... - [翻译此页]

27 Jun 2008 ... The skin is the largest organ of the body, it covers an area of 21 square feet (two square meters) and weighs around eleven pounds (five ...

www.articlesbase.com › Health › Acne - 网页快照 - 类似结果

Most Powerful Tea : Mate Coca | GEEK! - [翻译此页]

6 Sep 2009 ... 153624 0 Most Powerful Tea : Mate Coca. Mate coca, or with it's common name Coca tea, is made from the leaves of Coca plant. ...

www.turkgeek.net/most-**powerful-tea**-mate-coca/ - 网页快照

Fourth of July: 'I am America' Powerful tea party fave video debut ... - [翻译此页]

3 Jul 2010 ... This is a Tea Party favorite. Krista Branch debuts the video to her recent single release, "I Am America!" with powerful and inspiring ...

www.examiner.com/.../fourth-of-july-i-am-america-**powerful-tea**-party-fave-video-debut-by-krista-branch - 网页快照

Powerful TEA Party commercial - Target: Freedom - [翻译此页]

21 Aug 2009 ... and from a high school student yet. This resulted from a Mom in Alabama asking her high school son to help with a commercial for the Tea ...

5.1 Collocations

- Compositionality

  Compositional: if the meaning of the expression can be predicated from the meaning of the parts.

  Idiom: the most extreme example of non-compositionality

  kick the bucket

  an element of meaning added to the combination.

  strong tea (? *having great physical strength*)

- collocation, term, technical term, terminological phrase
- applications:

  NLG (*powerful tea, take a decision)*

  Computational lexicography, parsing, IR

5.2 搭配自动抽取：Frequency

Frequency

Four months of the New York Times Newswire: 14 million words

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

5.2 搭配自动抽取：Frequency

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

5.2 搭配自动抽取：Frequency

w	$C(strong, w)$	w	$C(powerful, w)$
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5
challenges	11	forces	5
challenge	11	chip	5
case	11	Germany	5
supporter	10	senators	4
signal	9	neighbor	4
man	9	magnet	4

5.3 搭配自动抽取：Mean and Variance

Mean and Variance

a. she knocked on his door

b. they knocked at the door

c. 100 women knocked on Donaldson's door

d. a man knocked on the metal front door

? hit, beat, rap

5.3 搭配自动抽取：Mean and Variance

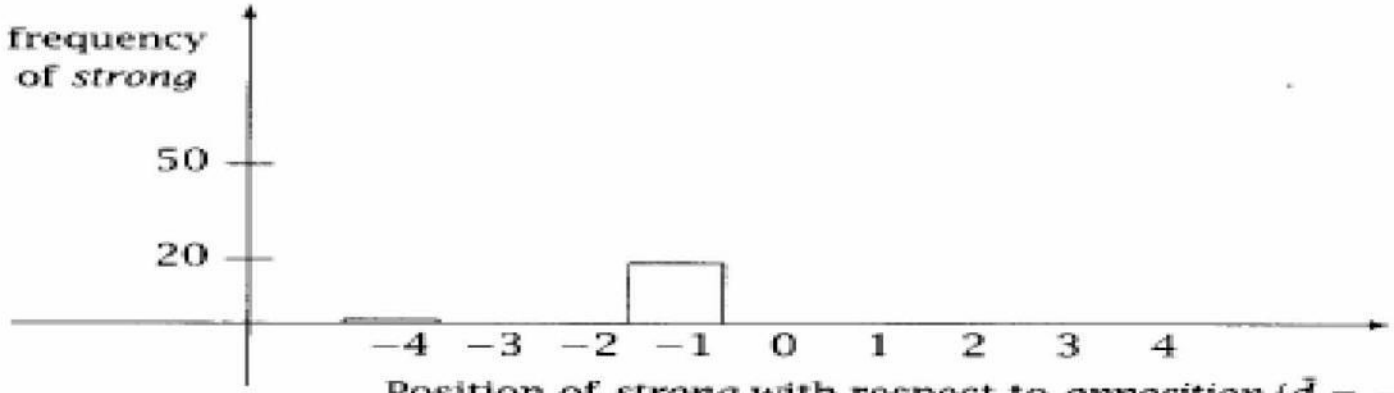
- To compute the mean and variance of the offset between the two words in the corpus
- focal word: *knocked*

$$\frac{1}{4}(3+3+5+5)=4.0$$

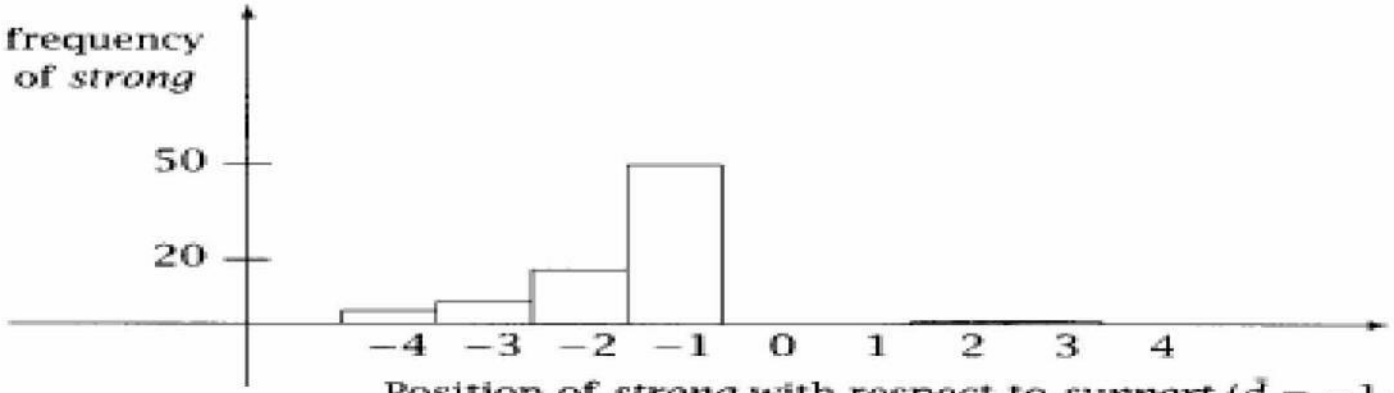
The door that he knocked on. -3

$$s^2=\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}$$

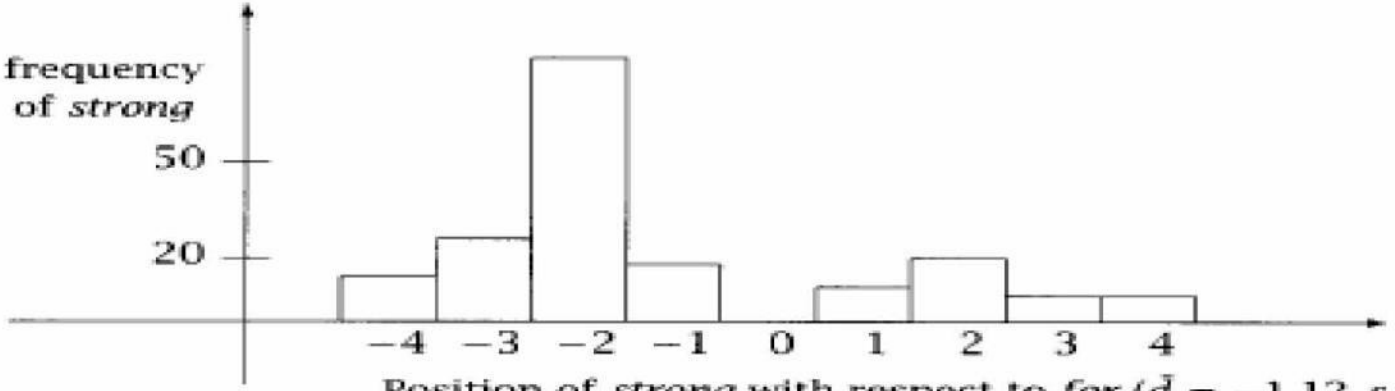
$$s=\sqrt{\frac{1}{3}((3-4.0)^2+(3-4.0)^2+(5-4.0)^2+(5-4.0)^2)}\approx 1.15$$



Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

teresting phrase. The pair *previous / games* (distance 2) corresponds to phrases like *in the previous 10 games* or *in the previous 15 games*; *minus / points* corresponds to phrases like *minus 2 percentage points*, *minus 3 percentage points* etc; *hundreds / dollars* corresponds to *hundreds of billions of dollars* and *hundreds of millions of dollars*.

Retrieving Collocations from Text: Xtract

Frank Smadja*
Columbia University

Natural languages are full of collocations, recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages. Recent work in lexicography indicates that collocations are pervasive in English; apparently, they are common

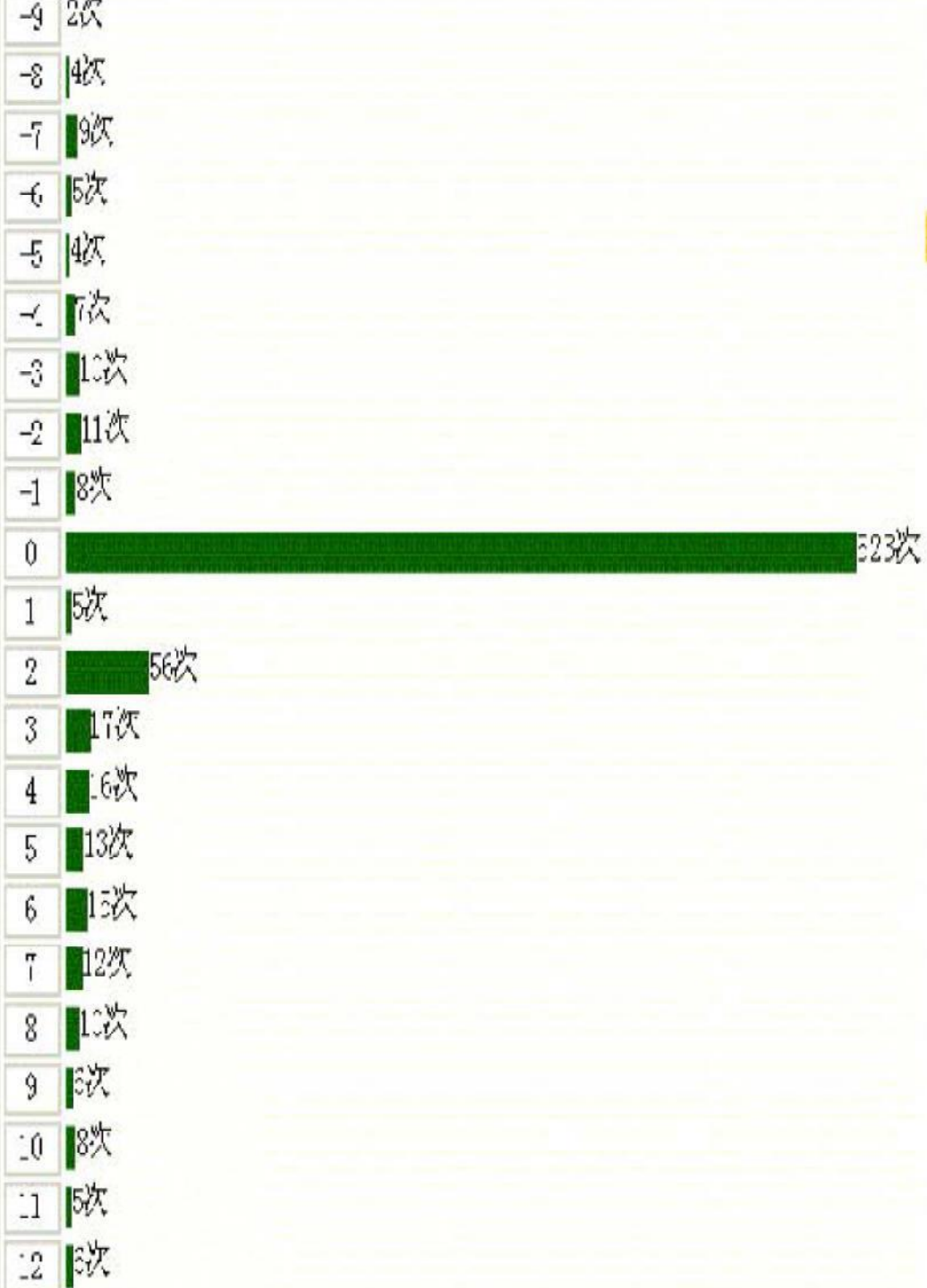
中国语文 1997 年第 1 期(总第 256 期)

汉语搭配定量分析初探*

孙茂松　黄昌宁　方　捷

提要　搭配在语言教学和语言信息处理中具有一定的应用价值。汉语搭配的研究仍停留在主要以人的主观判断为标准的定性分析阶段,缺乏定量数据的支待。本文借鉴了国外在语言学和语料库语言学两个方面关于搭配的研究成果,提出了包括强度、离散度及尖峰三项统计指标在内的搭配定量评估体系,构造了相应的搭配判断算法。作为对算法的初步测试,我们以一个约 710 万词次的新华社新闻语料库为工作平台,利用计算机对"能力"一词可能构成的搭配进行了全面分析。实验结果显示,就该词而言,算法自动发现搭配的准确率约为 33.94%。本项研究可望为语言学家客观、系统、一致地分析搭配提供定量辅助手段。

w	w_i	*Freq*	p_{-5}	p_{-4}	p_{-3}	p_{-2}	p_{-1}	p_1	p_2	p_3	p_4	p_5
takeover	possible	178	0	13	4	23	138	0	0	0	0	0
takeover	corporate	93	2	2	2	1	63	3	2	9	4	5
takeover	unsolicited	83	5	30	5	0	42	0	0	1	0	0
takeover	several	81	2	6	6	6	45	0	0	12	0	4
takeover	recent	76	5	4	6	5	17	0	0	36	2	1
takeover	new	75	4	3	6	28	27	0	1	4	2	0
takeover	unwanted	53	5	0	0	2	46	0	0	0	0	0
takeover	expensive	52	1	0	0	0	2	0	23	23	3	0
takeover	potential	50	1	0	1	3	42	0	0	0	2	1
takeover	big	47	0	0	0	4	15	0	0	5	21	2
takeover	friendly	41	0	3	3	1	25	0	0	2	3	4
takeover	unsuccessful	40	0	1	5	6	27	0	0	0	0	1
takeover	biggest	35	1	2	1	4	20	0	0	0	5	2
takeover	largest	32	0	1	3	20	3	0	0	0	0	5
takeover	old	28	0	8	6	0	14	0	0	0	0	0
takeover	unfriendly	26	0	0	0	0	18	0	0	0	0	8
takeover	rival	26	0	1	3	0	3	0	8	5	5	1
takeover	inadequate	26	5	10	2	0	0	0	0	9	0	0
takeover	initial	25	0	6	0	0	13	0	0	4	0	2
takeover	unwelcome	24	4	0	0	0	20	0	0	0	0	0
takeover	previous	24	0	2	0	4	18	0	0	0	0	0
takeover	federal	22	4	2	2	0	0	0	2	2	8	2
takeover	bitter	22	0	0	0	7	14	0	0	0	1	0
takeover	strong	19	0	4	3	5	4	0	0	1	0	2
takeover	hostile	16	0	6	0	0	10	0	0	0	0	0
takeover	attractive	16	1	0	5	3	7	0	0	0	0	0
takeover	unfair	13	0	0	0	0	13	0	0	0	0	0



Compound

网络(network) vs. 犯罪(crime)

Observation:

Total f: 907

peak(0)

网络 is likely to co-occur exactly before 犯罪 in S compound word

网络犯罪手段也不断更新

吞吐(import and export) vs. 能力(ability)

Observation：

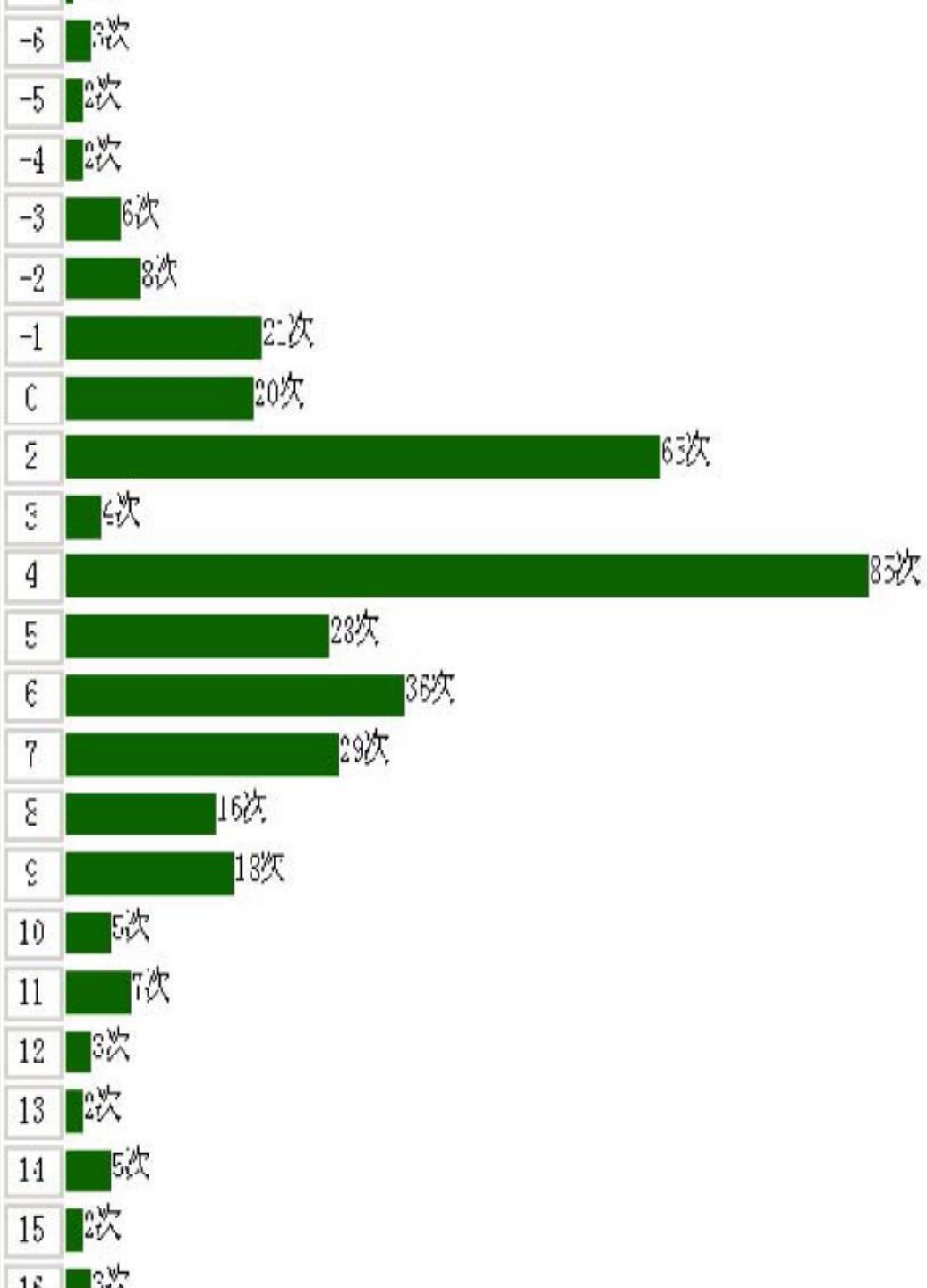
Total f: 240

(1) 吞吐 co-occurs exactly before 能力 in S

港口将新增吞吐能力4870万吨

西北地区最大铁路口岸吞吐货物能力达1000万吨

但箱运能力和吞吐量的



提高(raise) vs.
能力(ability)

Observation:

Total f: 398

(1) 能力 is likely to occur after 提高 in S

(2) location:

peak(-1)=21

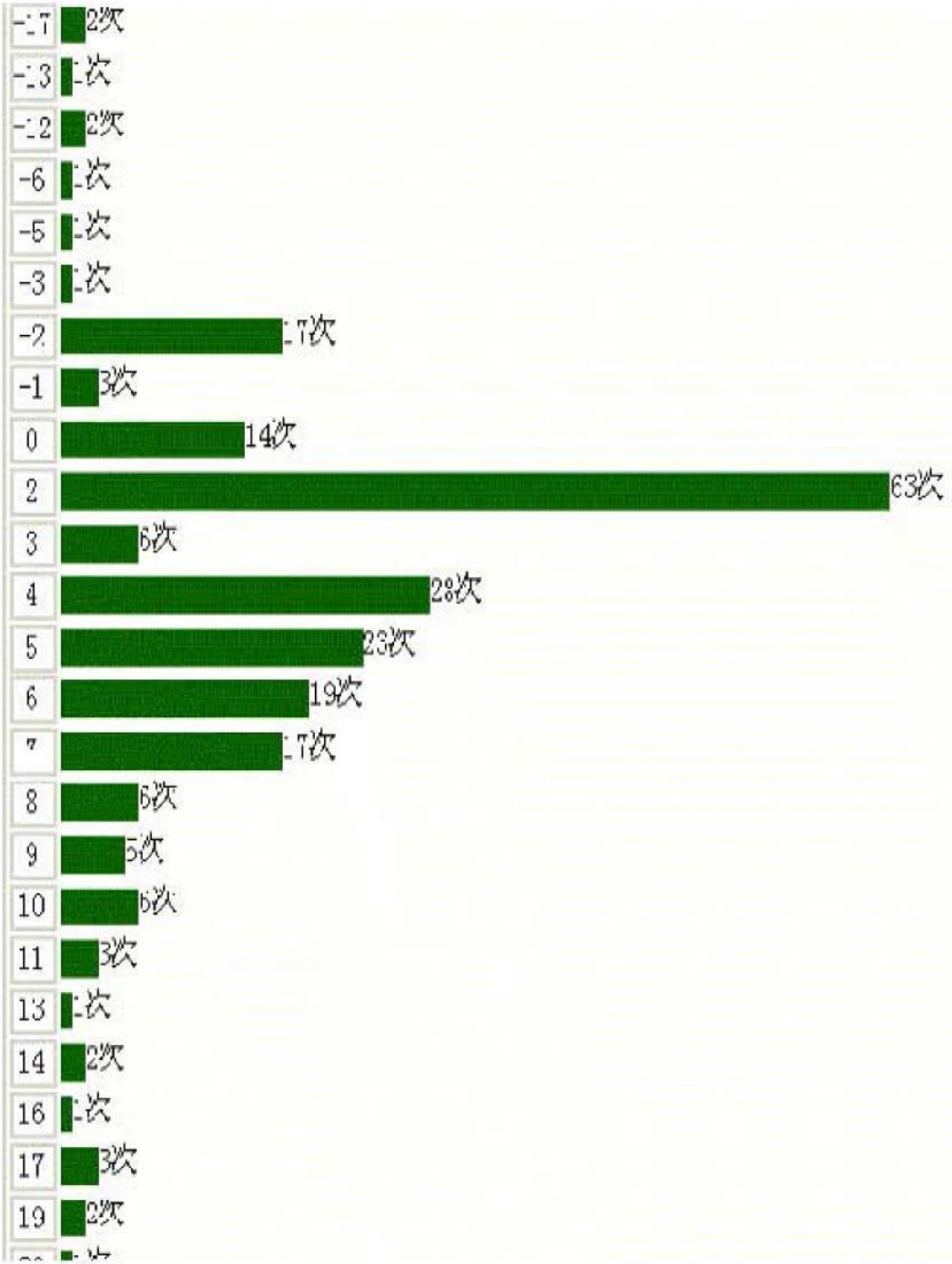
能力的提高

peak(4)=85

提高风险控制能力

peak(2)=63

如何提高写作能力



增强(strengthen) vs. 能力(ability)

Observation:

total f: 219

(1) 能力 is likely to occur after 增强 in S

(2) location:

peak(-2)=17

调控能力明显增强

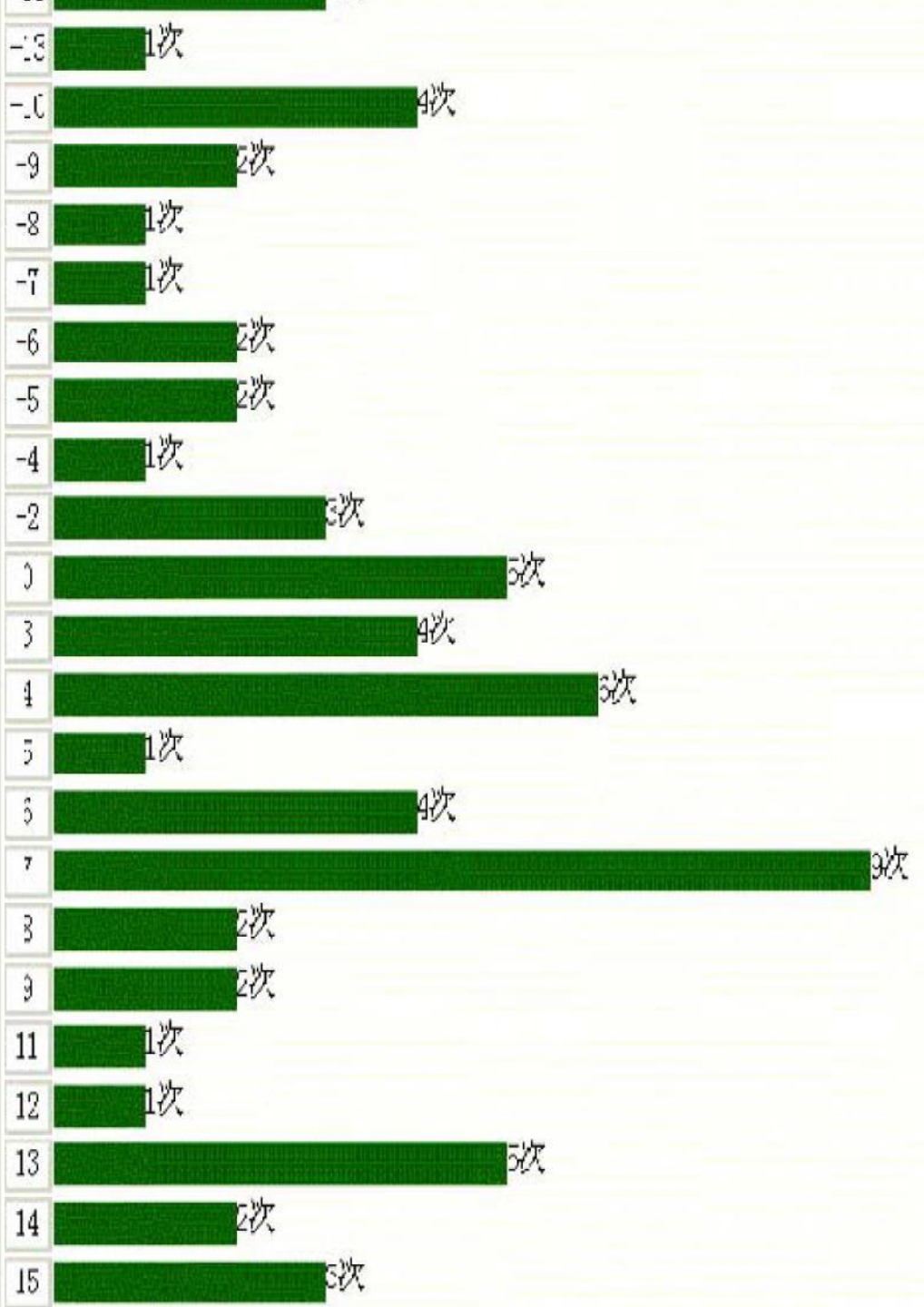
peak(2)=63

增强应变能力

peak(4)=28

增强调控监管能力

(3) collocation



增大(enlarge) vs.
能力(ability)

Observation:

total f: 74

(1) 能力 is likely to occur after 增大 in S

(2) location:
no obvious peak

(3) collocation ?

(a)

培养	判断	鉴赏	生产	竞争	制造	运输	加工	支付	偿还	平衡	消化	吸收
繁殖	实际	具备	缺乏	提高	强	弱	大	差	有限	增强	具有	业务
劳动	适应	领导	组织	分析	保护	发挥	工作	技术	专业	管理	创造	运输
发电	丧失	防御	装卸	指挥								

(b)

反应	扩大	综合	形成	达到	设计	抗灾	开采	影响	排水	客运	保障	承受
一定	执政	反应	安置	配套	不足	超过	出口	自力	创汇	动手	吞吐	增加
运行	足够	防务	操作	处理	作战	通信	同等	自给	自理	防守	减弱	现有
约束	作业	防卫	鉴别	通航	负重	不够	生存	隐蔽	科研	失去	抗病	炼油
腐蚀	后续	识别	抗旱	削弱	限制	识字	存储	自主	对抗	核算	机动	消费
分流	超出	防洪	自卫	干扰	免疫	再生	信任	过剩	供给	应急	饲养	运算
扑救	防疫	驾驭	筛选	参政	相应	采油	整体	通行	核定	载荷	维修	运载
接待	保存	分辨	保鲜	装备	耐寒	通车	转换	防范	自救	联运	决策	独到
起重	输送	新	有	开发	服务	群众	发展	测量	显示	突破	依靠	强化
控制	经营	供应	下降	监督	低	核	拥有					

5.4 搭配自动抽取：假设检验

Hypothesis Testing

If two words occur together more often than chance?

A null hypothesis H_0: no association between the words beyond chance occurrences

$$P(w^1 w^2) = P(w^1)P(w^2)$$

Significance level p =0.05, 0.01, 0.005

备择假设

5.4 搭配自动抽取：假设检验

Hypothesis Testing (The *t* test)

we compute the *t* statistic:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

$$\frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim t(n-1)$$

Here's an example of applying the *t* test. Our null hypothesis is that the mean height of a population of men is 158cm. We are given a sample of 200 men with $\bar{x} = 169$ and $s^2 = 2600$ and want to know whether this sample is from the general population (the null hypothesis) or whether it is from a different population of smaller men. This gives us the following *t* according to the above formula:

$$t = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$$

If you look up the value of *t* that corresponds to a confidence level of $\alpha = 0.005$, you will find 2.576.[4] Since the *t* we got is larger than 2.576, we can reject the null hypothesis with 99.5% confidence. So we can say that the sample is not drawn from a population with mean 158cm, and our probability of error is less than 0.5%.

$$P(new) = \frac{15828}{14307668} \qquad H_0 : P(new\ companies) = P(new)P(companies)$$

$$P(companies) = \frac{4675}{14307668} \qquad = \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

If the null hypothesis is true, then the process of randomly generating bigrams of words and assigning 1 to the outcome *new companies* and 0 to any other outcome is in effect a Bernoulli trial with $p = 3.615 \times 10^{-7}$ for the probability of *new company* turning up. The mean for this distribution is $\mu = 3.615 \times 10^{-7}$ and the variance is $\sigma^2 = p(1 - p)$ (see section 2.1.9), which is approximately p. The approximation $\sigma^2 = p(1 - p) \approx p$ holds since for most bigrams p is small.

The family of binomial distributions gives the number r of successes out of n trials given that the probability of success in any trial is p:

$$(2.13)\quad b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r} \quad \text{where} \quad \binom{n}{r} = \frac{n!}{(n - r)!r!} \quad 0 \le r \le n$$

The term $\binom{n}{r}$ counts the number of different possibilities for choosing r objects out of n, not considering the order in which they are chosen. Examples of some binomial distributions are shown in figure 2.3. The binomial distribution has an expectation of np and a variance of $np(1 - p)$.

It turns out that there are actually 8 occurrences of *new companies* among the 14,307,668 bigrams in our corpus. So, for the sample, we have that the sample mean is: $\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$. Now we have everything we need to apply the t test:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

5.4 搭配自动抽取：假设检验

Hypothesis Testing (of differences)

t	$C(w)$	$C(strong\ w)$	$C(powerful\ w)$	Word
3.1622	933	0	10	computers
2.8284	2337	0	8	computer
2.4494	289	0	6	symbol
2.4494	588	0	6	machines
2.2360	2266	0	5	Germany
2.2360	3745	0	5	nation
2.2360	395	0	5	chip
2.1828	3418	4	13	force
2.0000	1403	0	4	friends
2.0000	267	0	4	neighbor
7.0710	3685	50	0	support
6.3257	3616	58	7	enough
4.6904	986	22	0	safety
4.5825	3741	21	0	sales
4.0249	1093	19	1	opposition
3.9000	802	18	1	showing
3.9000	1641	18	1	sense
3.7416	2501	14	0	defense
3.6055	851	13	0	gains
3.6055	832	13	0	criticism

5.4 搭配自动抽取：假设检验

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\frac{(\bar{Y} - \bar{X}) - (\mu_2 - \mu_1)}{\left(S_1^2/n_1 + S_2^2/n_2\right)^{\frac{1}{2}}} \sim N(0,1)$$

$$t \approx \frac{P(v^1 w) - P(v^2 w)}{\sqrt{\frac{P(v^1 w) + P(v^2 w)}{N}}}$$

$$t \approx \frac{\frac{C(v^1 w)}{N} - \frac{C(v^2 w)}{N}}{\sqrt{\frac{C(v^1 w) + C(v^2 w)}{N^2}}}$$

$$= \frac{C(v^1 w) - C(v^2 w)}{\sqrt{C(v^1 w) + C(v^2 w)}}$$

5.4 搭配自动抽取：假设检验

Hypothesis Testing (Pearson's chi-square test)

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\frac{8+4667}{N} \times \frac{8+15820}{N} \times N \approx 5.2$$

	w_1 = *new*	$w_1 \neq$ *new*
w_2 = *companies*	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq$ *companies*	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

5.4 搭配自动抽取：假设检验

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

This formula gives the following χ^2 value for table 5.8:

$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8 + 4667)(8 + 15820)(4667 + 14287181)(15820 + 14287181)} \approx 1.55$$

Looking up the χ^2 distribution in the appendix, we find that at a probability level of $\alpha = 0.05$ the critical value is $\chi^2 = 3.841$ (the statistic has one degree of freedom for a 2-by-2 table). So we cannot reject the null hypothesis that *new* and *companies* occur independently of each other. Thus *new companies* is not a good candidate for a collocation.

5.4 搭配自动抽取：假设检验

	cow	¬ *cow*
vache	59	6
¬ *vache*	8	570934

$$\chi^2 = 456400.$$

5.4 搭配自动抽取：假设检验

Hypothesis Testing (Likelihood ratios)

- **Hypothesis 1.** $P(w^2|w^1) = p = P(w^2|\neg w^1)$
- **Hypothesis 2.** $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

Hypothesis 1 is a formalization of independence (the occurrence of w^2 is independent of the previous occurrence of w^1), Hypothesis 2 is a formalization of dependence which is good evidence for an interesting collocation.[6]

We use the usual maximum likelihood estimates for p, p_1 and p_2 and write c_1, c_2, and c_{12} for the number of occurrences of w^1, w^2 and w^1w^2 in the corpus:

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

5.4 搭配自动抽取：假设检验

- **Hypothesis 1.** $P(w^2|w^1) = p = P(w^2|\neg w^1)$
- **Hypothesis 2.** $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

the likelihood of getting the counts for w^1, w^2 and w^1w^2 that we actually observed is then $L(H_1) = b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)$ for Hypothesis 1 and $L(H_2) = b(c_{12}; c_1, p_1)b(c_2 - c_{12}; N - c_1, p_2)$ for Hypothesis 2. Ta-

	H_1	H_2
$P(w^2\|w^1)$	$p = \frac{c_2}{N}$	$p_1 = \frac{c_{12}}{c_1}$
$P(w^2\|\neg w^1)$	$p = \frac{c_2}{N}$	$p_2 = \frac{c_2 - c_{12}}{N - c_1}$
c_{12} out of c_1 bigrams are w^1w^2	$b(c_{12}; c_1, p)$	$b(c_{12}; c_1, p_1)$
$c_2 - c_{12}$ out of $N - c_1$ bigrams are $\neg w^1w^2$	$b(c_2 - c_{12}; N - c_1, p)$	$b(c_2 - c_{12}; N - c_1, p_2)$

Table 5.11 How to compute Dunning's likelihood ratio test. For example, the likelihood of hypothesis H_2 is the product of the last two lines in the rightmost

5.4 搭配自动抽取：假设检验

The log of the likelihood ratio λ is then as follows:

$$\begin{aligned}
\log\lambda &= \log\frac{L(H_1)}{L(H_2)} \\
&= \log\frac{\mathrm{b}(c_{12},c_1,p)\mathrm{b}(c_2-c_{12},N-c_1,p)}{\mathrm{b}(c_{12},c_1,p_1)\mathrm{b}(c_2-c_{12},N-c_1,p_2)} \\
&= \log L(c_{12},c_1,p)+\log L(c_2-c_{12},N-c_1,p) \\
&\quad -\log L(c_{12},c_1,p_1)-\log L(c_2-c_{12},N-c_1,p_2)
\end{aligned}$$

where $L(k,n,x)=x^k(1-x)^{n-k}$.

$-2 \log \lambda$	$C(w^1)$	$C(w^2)$	$C(w^1 w^2)$	w^1	w^2
1291.42	12593	932	150	most	powerful
99.31	379	932	10	politically	powerful
82.96	932	934	10	powerful	computers
80.39	932	3424	13	powerful	force
57.27	932	291	6	powerful	symbol
51.66	932	40	4	powerful	lobbies
51.52	171	932	5	economically	powerful
51.05	932	43	4	powerful	magnet
50.83	4458	932	10	less	powerful
50.75	6252	932	11	very	powerful
49.36	932	2064	8	powerful	position
48.78	932	591	6	powerful	machines
47.42	932	2339	8	powerful	computer
43.23	932	16	3	powerful	magnets
43.10	932	396	5	powerful	chip
40.45	932	3694	8	powerful	men
36.36	932	47	3	powerful	486
36.15	932	268	4	powerful	neighbor
35.24	932	5245	8	powerful	political
34.15	932	3	2	powerful	cudgels

Table 5.12 Bigrams of *powerful* with the highest scores according to Dunning's likelihood ratio test.

5.5 搭配自动抽取：相对频度

(Relative frequency ratios)

Table 5.13 shows ten bigrams that occur exactly twice in our referen corpus (the 1990 *New York Times* corpus). The bigrams are ranked a cording to the ratio of their relative frequencies in our 1990 referen corpus versus their frequencies in a 1989 corpus (again drawn from t months August through November). For example, *Karim Obeid* occurs times in the 1989 corpus. So the relative frequency ratio r is:

$$r = \frac{\frac{2}{14307668}}{\frac{68}{11731564}} \approx 0.024116$$

Ratio	1990	1989	w^1	w^2
0.0241	2	68	Karim	Obeid
0.0372	2	44	East	Berliners
0.0372	2	44	Miss	Manners
0.0399	2	41	17	earthquake
0.0409	2	40	HUD	officials
0.0482	2	34	EAST	GERMANS
0.0496	2	33	Muslim	cleric
0.0496	2	33	John	Le
0.0512	2	32	Prague	Spring
0.0529	2	31	Among	individual

Table 5.13 Damerau's frequency ratio test. Ten bigrams that occurred twice in the 1990 *New York Times* corpus, ranked according to the (inverted) ratio of relative frequencies in 1989 and 1990.

5.6 搭配自动抽取：点对互信息

Mutual Information

$$I(x', y') = \log_2 \frac{P(x'y')}{P(x')P(y')} = \log_2 \frac{P(x'|y')}{P(x')} = \log_2 \frac{P(y'|x')}{P(y')}$$

$$I(X;\ Y) = E_{p(x,y)} \log \frac{p(X,Y)}{p(X)\,p(Y)}$$

Symbol	Definition	Current use	Fano
$I(x, y)$	$\log \frac{p(x,y)}{p(x)\,p(y)}$	pointwise mutual information	mutual information
$I(X;\ Y)$	$E \log \frac{p(X,Y)}{p(X)\,p(Y)}$	mutual information	average MI/expectation of MI

Table 5.17 Different definitions of *mutual information* in (Cover and Thomas 1991) and (Fano 1961).

5.6 搭配自动抽取：点对互信息

发展	35264	7.3850
国家	28022	5.3231
市场	25269	7.7532
建设	21232	7.8517
问题	21142	8.9182
记者	20944	8.3432
我们	19912	7.1875
全国	18137	4.9564
改革	16881	9.0339
国际	15600	6.0306
进行	15121	6.2356
社会	14493	5.7883
老人	1000	3.6251
有着	999	3.0236
协商	998	5.9236
下来	994	3.3679
工会	989	1.3247
婚礼	100	9.6758
认清	100	3.7308
有功	100	1.0538

5.6 搭配自动抽取：点对互信息

鳏鳇	1	24.2707
妯娌	2	23.2707
糌粑	5	21.1008
馄饨	9	20.9488
吝啬	12	19.1552
镶嵌	37	17.0814
坍塌	19	15.9902
签署	1779	11.0060
基层	1848	8.0001
唱片	166	7.9987
相当	1328	6.0020
壮大	264	5.0013
人权	1603	3.9969
看上	84	0.7987
在家	317	-0.5878
后事	15	-2.0107
发报	24	-3.0030
指法	2	-4.0004
不第	1	-7.0854

5.6 搭配自动抽取：点对互信息

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

Table 5.14 Finding collocations: Ten bigrams that occur with frequency 20, ranked according to mutual information.

of the other). For perfect dependence we have:

$$I(x, y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)}{P(x)P(y)} = \log \frac{1}{P(y)}$$

That is, among perfectly dependent bigrams, as they get rarer, their mutual information *increases*.

For perfect independence we have:

$$I(x, y) = \log \frac{P(xy)}{P(x)P(y)} = \log \frac{P(x)P(y)}{P(x)P(y)} = \log 1 = 0$$

I_{1000}	w^1	w^2	w^1w^2	Bigram	I_{23000}	w^1	w^2	w^1w^2	Bigram
16.95	5	1	1	Schwartz eschews	14.46	106	6	1	Schwartz eschews
15.02	1	19	1	fewest visits	13.06	76	22	1	FIND GARDEN
13.78	5	9	1	FIND GARDEN	11.25	22	267	1	fewest visits
12.00	5	31	1	Indonesian pieces	8.97	43	663	1	Indonesian pieces
9.82	26	27	1	Reds survived	8.04	170	1917	6	marijuana growing
9.21	13	82	1	marijuana growing	5.73	15828	51	3	new converts
7.37	24	159	1	doubt whether	5.26	680	3846	7	doubt whether
6.68	687	9	1	new converts	4.76	739	713	1	Reds survived
6.00	661	15	1	like offensive	1.95	3549	6276	6	must think
3.81	159	283	1	must think	0.41	14093	762	1	like offensive

Table 5.16 Problems for Mutual Information from data sparseness. The table shows ten bigrams that occurred once in the first 1000 documents in the reference corpus ranked according to mutual information score in the first 1000 documents (left half of the table) and ranked according to mutual information score in the entire corpus (right half of the table). These examples illustrate that a large proportion of bigrams are not well characterized by corpus data (even for large corpora) and that mutual information is particularly sensitive to estimates that are inaccurate due to sparseness.

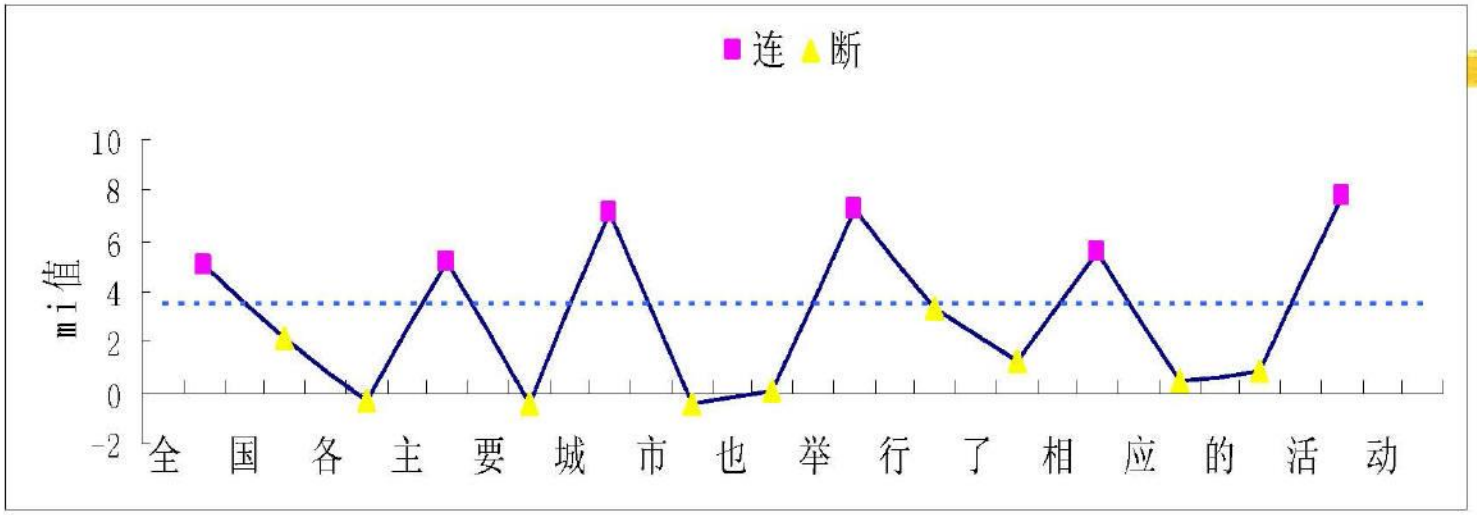
5.6 搭配自动抽取：点对互信息

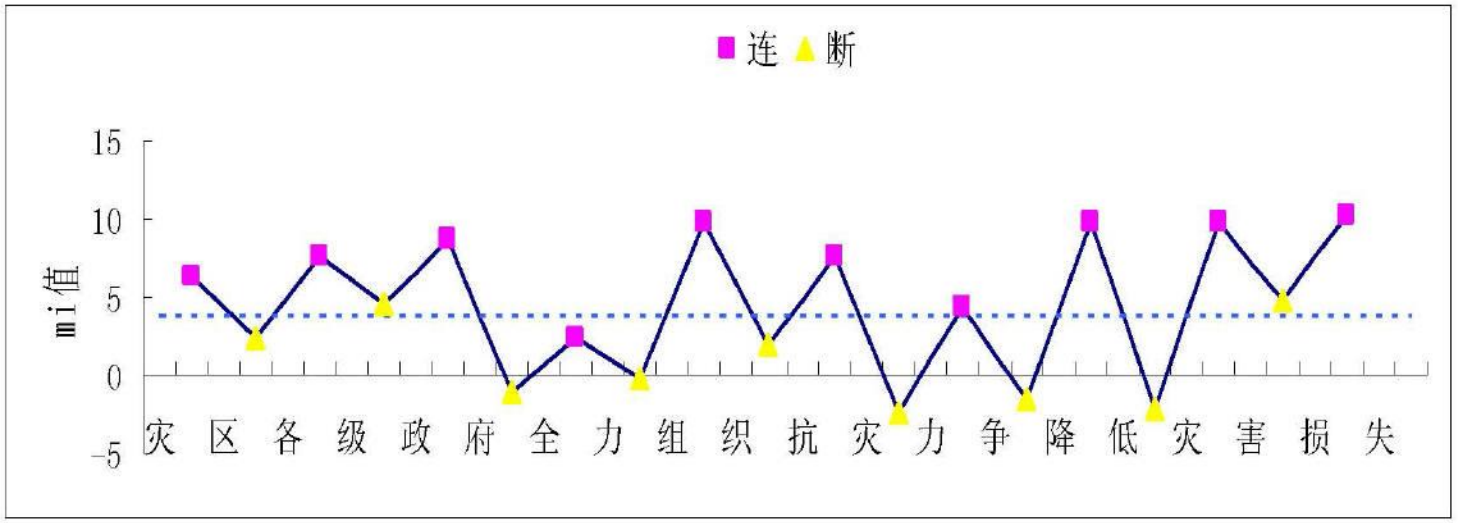
Frequency f_{xy}

Mutual Information $MI(x,y) = \log_2 \frac{P(y|x)}{P(y)} = \log_2 \frac{P_{xy}}{P_x P_y}$

Frequency*Mutual Information $PMI(x,y) = P(x,y)MI(x,y)$

5.7 Collocations: Dealing with unknown words





5.7 Collocations: Dealing with unknown words

定义. 对汉字串xyz，汉字y相对于x及z的t-测试定义为：

$$t_{x,z}(y)=\frac{p(z\mid y)-p(y\mid x)}{\sqrt{\sigma^2(p(z\mid y))+\sigma^2(p(y\mid x))}}$$

其中$p(y\mid x)$、$p(z\mid y)$分别是y关于x、z关于y的条件概率，$\sigma^2(p(y\mid x))$、$\sigma^2(p(z\mid y))$是各自的方差。

从 t-测试的定义，可知：

(1) 如果$t_{x,z}(y)>0$，则y倾向于与其后继字z连（亦即倾向于与其前趋字x断）。且值越大，倾向越强；

(2) 如果$t_{x,z}(y)<0$，则y倾向于与其前趋字x连（亦即倾向于与其后继字z断）。且绝对值越大，倾向越强；

(3) 如果$t_{x,z}(y)=0$，则无任何倾向。

5.7 Collocations: Dealing with unknown words

定义. 对汉字串$vxyw$，汉字 x、y 之间的 t-测试差（或称为汉字 x、y 间位置的 t-测试差）定义为：

$$dts(x,y) = t_{v,y}(x) - t_{x,w}(y)$$

显然，t-测试差是附着于字间位置的。$dts(x,y)$ 的作用可分情形讨论：

(1) $t_{v,y}(x) > 0$，且 $t_{x,w}(y) < 0$：

此时 x、y 相互吸引，必有 $dts(x,y) > 0$，x、y 之间倾向于连。

(2) $t_{v,y}(x) < 0$，且 $t_{x,w}(y) > 0$：

此时 x、y 相互排斥，必有 $dts(x,y) < 0$，x、y 之间倾向于断。

(3) $t_{v,y}(x) > 0$，且 $t_{x,w}(y) > 0$，或者 $t_{v,y}(x) < 0$，且 $t_{x,w}(y) < 0$：

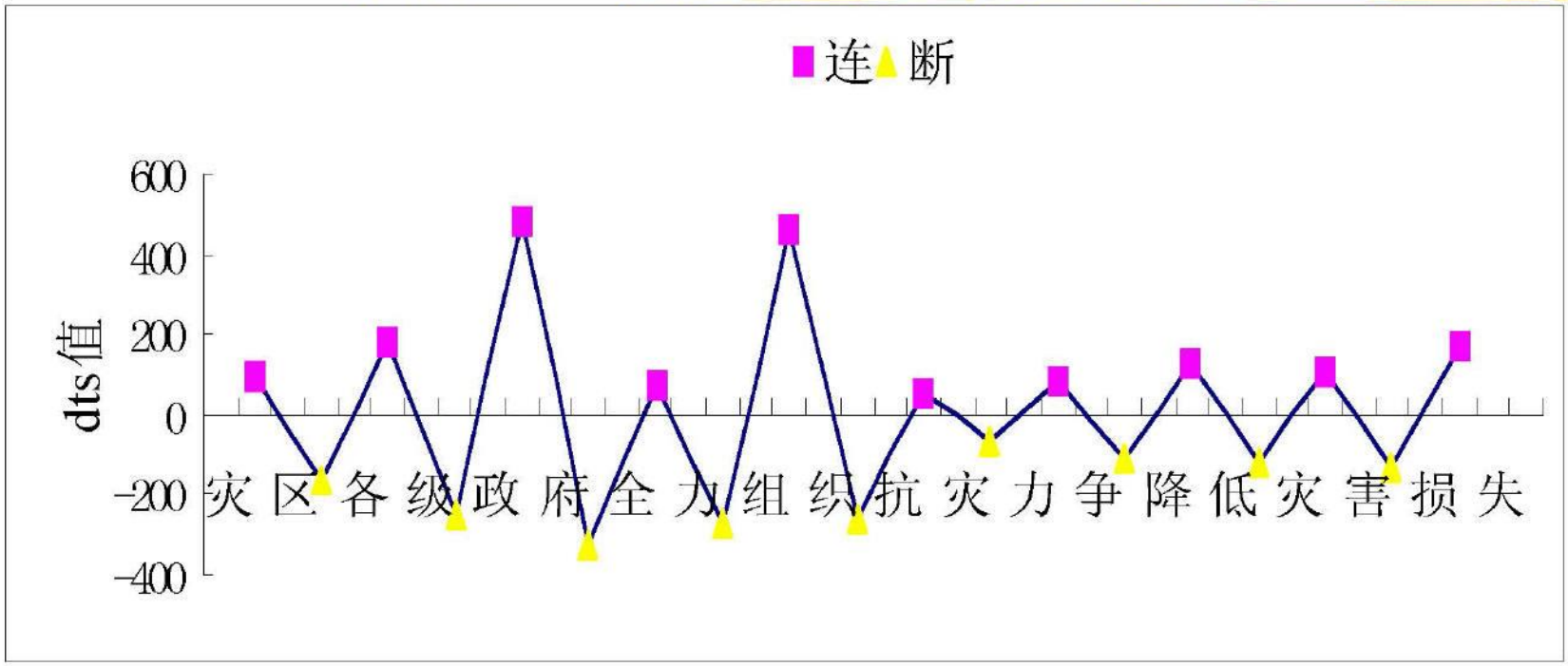
此时，将产生“竞争”，如果 $dts(x,y) > 0$，则竞争的结果倾向于连；

如果 $dts(x,y) < 0$，则竞争的结果倾向于断。

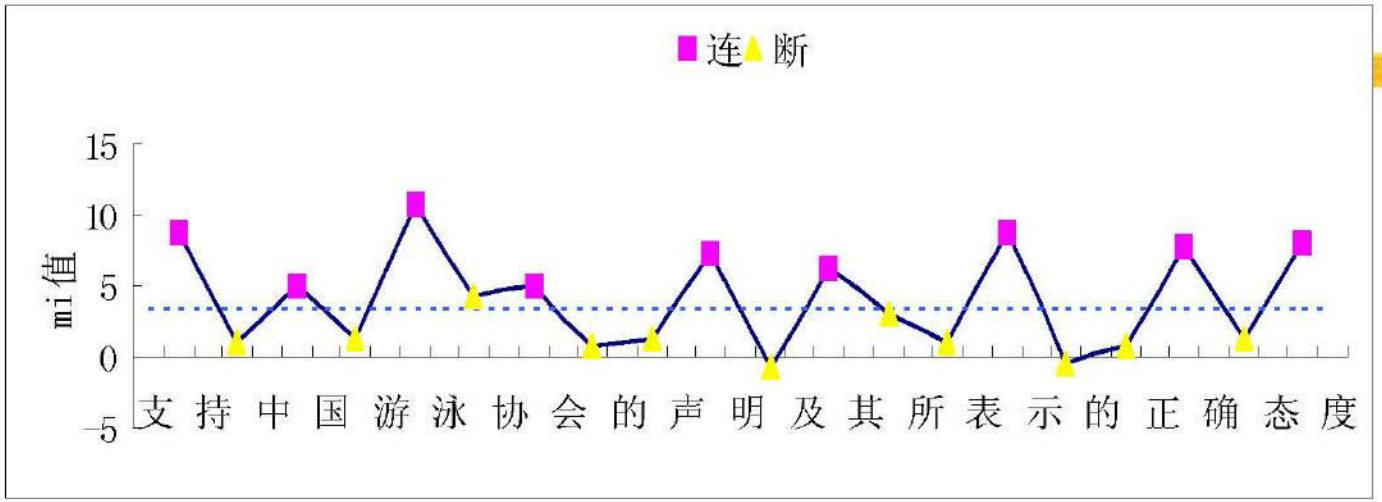
无论何种情形，均有：$dts(x,y)$ 越大，连的倾向越大，$dts(x,y)$ 值越小，断的倾向越大。$dts(x,y)$ 为 0 时，则无任何倾向（此为 t-测试差的“盲点”）。

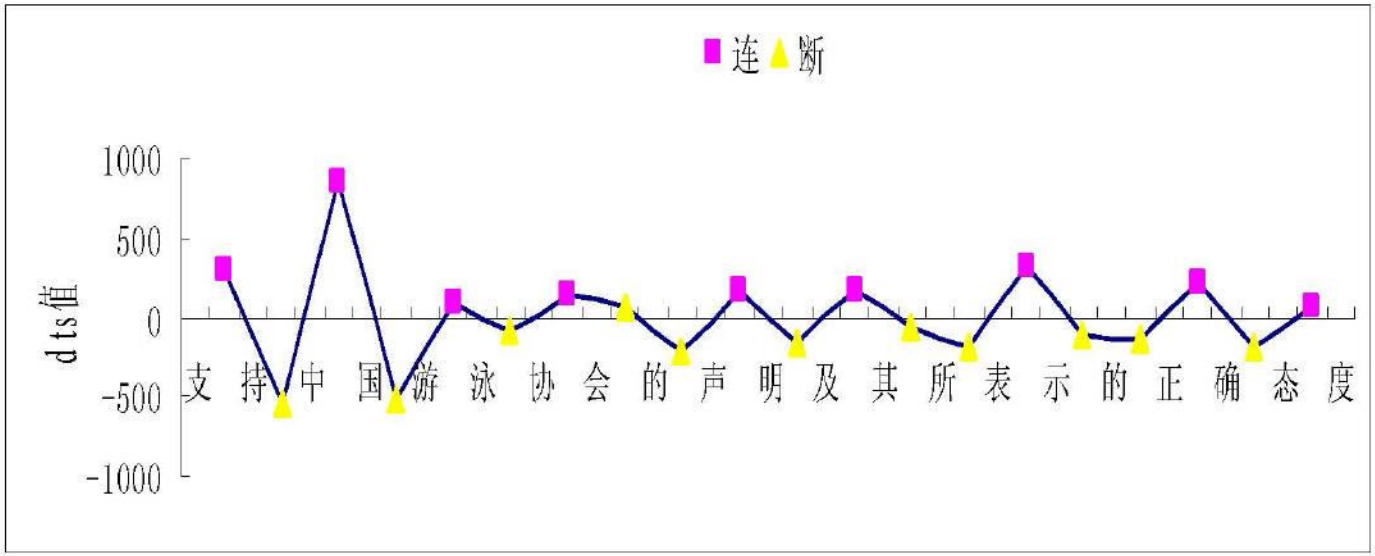
$mi(x,y)$ 与 $dts(x,y)$ 的主要差别在于：前者是 x、y 结合力的绝对度量，与上下文无关（仅涉及 2 个字）；前者则是 x、y 结合力的相对度量，与上下文有关（共涉及 4 个字）。

5.7 Collocations: Dealing with unknown words



5.7 Collocations: Dealing with unknown words





5.7 Collocations: Dealing with unknown words

注意到 mi 和 dts 变化范围相差甚远，所以需先进行标准化：

$$mi^*(x,y) = \frac{mi(x,y) - \mu_{mi}}{\sigma_{mi}}$$

$$dts^*(x,y) = \frac{dts(x,y) - \mu_{dts}}{\sigma_{dts}}$$

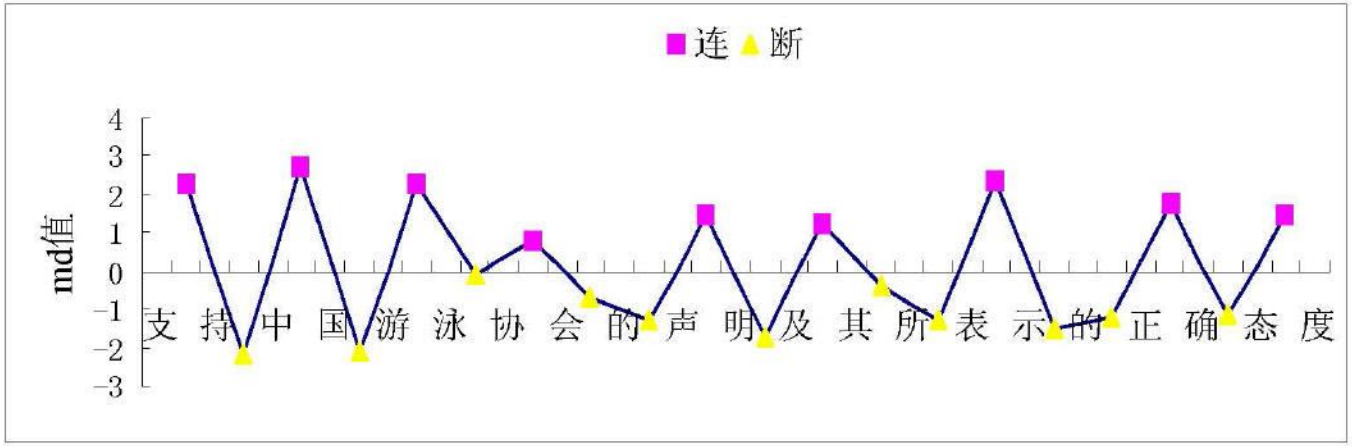
其中 μ_{mi}、μ_{dts}、σ_{mi}、σ_{dts} 分别为 mi 及 dts 的均值与均方差。并且有：μ_{dts} 的理论值为 0。μ_{mi}、σ_{mi}、σ_{dts} 的实验值依次为 3.50、3.48、217.80。

然后通过下式将互信息和 t-测试叠加起来：

$$md(x,y) = mi^*(x,y) + \lambda \times dts^*(x,y)$$

显然，md 均值 μ_{md} 的理论值为 0。md 均方差 σ_{md} 的实验值为 1.43。

经实验测定，λ 取 0.60 时分词效果最好。



5.7 Collocations: Dealing with unknown words