Collocations

สุธี สุดประเสริฐ

จากหนังสือ Foundations of Statistical Natural Langauge Processing บทที่ 5

Topics

- Introduction
- Frequency
- Mean and Variance
- Hypothesis Testing
- Mutual Information
- The Notion of Collocation

Introduction

- Collocation คือ การอธิบายที่ประกอบด้วยคำมากกว่าหนึ่งคำ ซึ่งการรวมของคำเหล่านั้นเป็นไปตามความความคุ้นเคยที่พูดตามๆ กันมา
  - strong tea
  - weapon of mass destruction
  - กาแฟแก่ (เหล้าแก่? น้ำหวานแก่? ไวน์แก่?)
  - ฝนตกหนัก (แดดออกหนัก? ลมพัดหนัก?)
- เป็นการยากที่จะให้เหตุผล แต่ผู้พูดในภาษานั้นรู้ได้เอง (จากความเคยชิน)

Introduction

- Compositional expression
  - สามารถเดาความหมายจากส่วนต่างๆ ได้
    - bad boys, strong men
- Collocation
  - สามารถเดาความหมายได้บางส่วน
    - strong tea, stiff breeze
- Idiom
  - ไม่สามารถเดาความหมายได้เลย
    - kick the bucket, go west

Introduction

- term, technical term and terminological phrase
  - มีความเกี่ยวเนื่องกันโดยตรงกับ collocation
  - คือผลรับจากการหา collocation ในข้อความเฉพาะทาง

Introduction

- ความสำคัญ
  - Natural lanugage generation
    - ไม่สร้างประโยคแปลกๆ เช่น powerful tea
  - Computaional lexicography
    - หา collocation ที่สำคัญ เพื่อเพิ่มเข้าในในพจนานุกรม
  - Parsing
    - ช่วยให้ parser วิเคราะห์ประโยคที่มี collocation
  - Corpus linguistic research
    - ศึกษาเรื่องแรงผลักดันทางวัฒนธรรมที่มีผลต่อภาษา

Frequency

- การนับความถี่ของคำที่เกิดติดกันบ่อยๆ เป็นวิธีที่ง่ายที่สุดในการหา collocations

C(w1, w2)	w1	w2
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

Frequency

- ปรับปรุงโดยใช้ POS มาช่วยคัดกรองเฉพาะคู่ที่น่าสนใจ

C(w1,w2)	w1	w2	Tags
11487	New	York	A N
7261	United	States	A N
3301	Los	Angeles	N N
3191	**last**	**year**	**A N**
2699	Saudi	Arabia	N N
2514	**last**	**week**	**A N**
2378	vice	president	A N
2161	Persian	Gulf	A N
2106	San	Francisco	N N
2001	Middle	East	A N

C(w1,w2)	w1	w2	Tags
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

Frequency

C(strong, w)			
support	50	criticism	13
safety	22	possibility	11
sales	21	feelings	11
opposition	19	demand	11
showing	18	challenges	11
sense	18	challenge	11
message	15	case	11
defense	14	supporter	10
gains	13	signal	9
evidence	13	**man**	9

C(powerful, w)			
force	13	weapons	5
computers	10	post	5
position	8	people	5
men	8	nation	5
computer	8	forces	5
man	7	chip	5
symbol	6	Germany	5
military	6	senators	4
machines	6	neighbor	4
country	6	magnet	4

Mean and Variance

- การนับความถี่ใช้ได้ดีกับวลีที่เกิดคงที่ (fixed-phrases) แต่ว่า collocation หลายตัวที่การเกิดแบบไม่คงที่ เช่น knock กับ door
  - she **knocked** on his **door**
  - they **knocked** at the **door**
  - 100 women **knocked** on Donaldson's **door**
  - a man **knocked** on the metal front **door**

Mean and Variance

- Mean: ค่าเฉลี่ยของระยะระหว่างคำสองคำ
  - she **knocked** on his **door** ($d_1 = 3$)
  - they **knocked** at the **door** ($d_2 = 3$)
  - 100 women **knocked** on Donaldson's **door** ($d_3 = 5$)
  - a man **knocked** on the metal front **door** ($d_4 = 5$)

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

Mean and Variance

- Variance: วัดความเบี่ยงเบนจากค่าเฉลี่ยของระยะทางระหว่างคำ

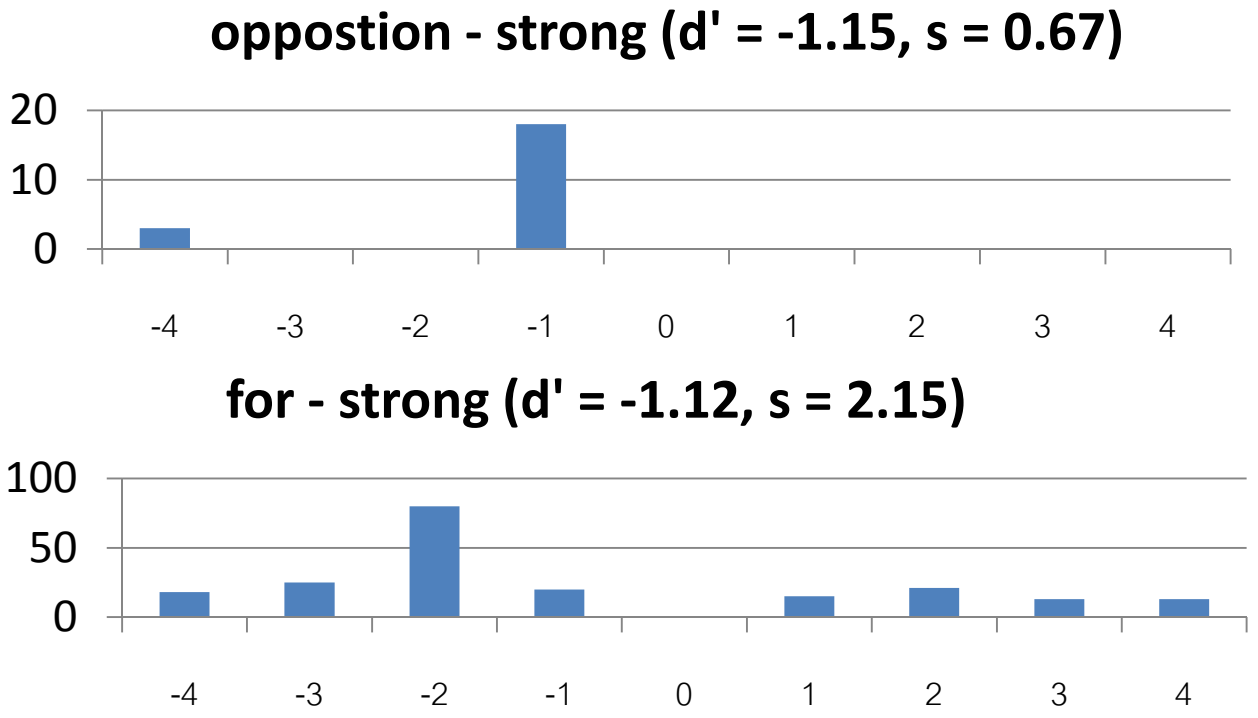
$$s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$$

- จากตัวอย่าง knock-door จะได้

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

Mean and Variance

- เราสามารถใช้ข้อสมมุติฐานว่าคู่ที่มีค่า variance ต่ำๆ น่าจะเป็น collocation



ใช้ window ขนาด 9 โดยคำที่สนใจอยู่ตรงกลาง (หน้า 4 หลัง 4)

Mean and Variance

s	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

New York

previous 10 games

minus 2 percentage points

hundreds of billions of dollars

strong {business} support

powerful {lobbying} organizations
powerful {tool for} organizations

Richard {M.} Nixon

Garrison said
said Garrison

Hypothesis Testing

- การที่คำคู่หนึ่งมี frequency และ variance ต่ำอาจจะเกิดจากความบังเอิญ โดยที่คำคู่นั้นไม่ได้เป็น collocation ก็ได้
  - เช่น new companies
- การหาว่าเหตุการณ์หนึ่งเกิดขึ้นโดยบังเอิญหรือไม่ เป็นปัญหาหนึ่งในวิชาสถิติ ซึ่งมักจะถูกเรียกว่า hypothesis testing
- สำหรับการตั้งสมมุติฐานว่าสองเหตุการณ์ไม่มีความเกี่ยวข้องกัน จะเรียกว่า null hypothesis หรือ H_0

Null Hypothesis

- ตัวอย่าง: ยาชนิดหนึ่งอาจมีผลต่อการลดโอกาสของการเป็นโรคหัวใจ ซึ่ง H_0 ที่เป็นไปได้คือ
  - ยาชนิดนี้ไม่ได้ลดโอกาสการเป็นโรคหัวใจ
  - ยาชนิดนี้ไม่มีผลกระทบต่อโอกาสการเป็นโรคหัวใจ
- การทดสอบ H_0 สามารถทำได้โดยการแบ่งผู้ป่วยเป็น 2 กลุ่ม คือ กลุ่มได้รับยา กับ กลุ่มไม่ได้รับยา ถ้ากลุ่มที่ให้ยามีการเปลี่ยนแปลงที่มีนัยสำคัญทางสถิติ (statistically significant) H_0 จะถูกปฏิเสธ

Null Hypothesis

- สำหรับการหา collocation เราต้องการที่จะทราบว่าคำสองคำเกิดขึ้นพร้อมกันโดยไม่ได้เกิดจากความบังเอิญ
- H_0 คือ คำคู่นั้นไม่มีความสัมพันธ์กันเกินไปกว่าการเกิดขึ้นโดยบังเอิญ
- ถ้าความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้นภายใต้ H_0 มีค่าต่ำ (น้อยกว่า 0.05, 0.01, 0.005 หรือ 0.001) แสดงว่า H_0 ถูกปฏิเสธ
- ถ้าคำสองคำไม่มีความสัมพันธ์ต่อกัน

$$P(w_1 w_2) = P(w_1)P(w_2)$$

The t test (Student's t-test)

- การทดสอบทางสถิติที่ใช้ในการหาความแตกต่างระหว่างค่าเฉลี่ยที่จากการสังเกต $\bar{x}$ กับค่าเฉลี่ยที่คาดคะเนไว้ μ

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

The t test

- ตัวอย่าง: H_0 คือ ความสูงเฉลี่ยของผู้ชายอาจเป็น 158 cm ถ้าเลือกกลุ่มตัวอย่างเป็นชาย 200 คน หาค่าความสูงเฉลี่ยได้ 169 cm ค่าความเบี่ยงเบน s^2 ได้ 2600

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

- ถ้าเทียบจากตารางของค่า t โดยดูที่ค่าระดับความสำคัญ 0.005 จะพบว่ามีค่า t เป็น 2.576 แต่ว่า t ที่หาได้มีค่ามากกว่า ดังนั้น ค่าระดับความสำคัญจึงต้องน้อยกว่า 0.005 ดังนั้น H_0 จึงไม่เป็นจริง

The t test

	p	0.05	0.025	0.01	**0.005**	0.001	0.0005
	C	90%	95%	98%	**99%**	99.8%	99.9%
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
(z)	∞	1.645	1.960	2.326	**2.576**	3.091	3.291

The t test

- ตัวอย่างการใช้ t-test ในการหา collocation เริ่มจากการหาค่า t ระหว่างคู่คำ เช่น new กับ companies
- ใช้ maximum likelihood ประมาณความน่าจะเป็นของคำ

$$P(new) = \frac{15828}{14307668}$$

$$P(companies) = \frac{4675}{14307668}$$

- สำรวจจากคลังประโยคขนาด 14307668 คำ นับ new ได้ 15828 คำ และนับ companies ได้ 4675 คำ

The t test

$$H_0 : P(new\ companies) = P(new)P(companies)$$
$$= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

- ถ้า H_0 เป็นจริง การเกิดคู่กันของ new และ companies จะเป็นแบบสุ่ม ถ้าให้ผลลัพธ์ของการคู่กันของสองคำนั้นเป็น 1 และสำหรับคู่อื่นให้เป็น 0 เราจะได้ Bernoulli trial ที่ $p = 3.615 \times 10^{-7}$ และ $s^2 = p(1-p) \approx p$

The t test

- ในคลังประโยค new และ companies เกิดคู่กัน 8 ครั้ง ดังนั้น

$$\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$$

- จากทั้งหมดเราจึงหาค่า t ได้

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} \approx 0.999932$$

The t test

- ค่า t ของ new และ companies น้อยกว่า 2.576 ซึ่ง ค่าความสำคัญเป็น 0.005 ดังนั้นเราจึงไม่สามารถปฏิเสธ H_0 ที่กล่าวว่า new และ companies ไม่ขึ้นต่อกัน

- สรุป new และ companies จึงไม่น่าเป็น collocation

The t test

t	C(w1)	C(w2)	C(w1 w2)	w1	w2
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

ตารางแสดงค่า t ของ bigram ที่เกิดขึ้นในคลังประโยค 20 ครั้ง

Hypothesis testing of differences

- t-test นอกจากจะใช้เพื่อหาคู่คำที่มีโอกาสเป็น collocation แล้ว ยังสามารถใช้ในการหาว่าคำสองคำมีความหมายแตกต่างกันอย่างไร ซึ่งมีประโยชน์ในการสร้างพจนานุกรม
  - strong และ powerful
- H_0 คือ ทั้งสองคำมีความหมายเหมือนกัน

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

Hypothesis testing of differences

t	C(w)	C(strong w)	C(powerful w)	w
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

Hypothesis testing of differences

t	C(w)	C(strong w)	C(powerful w)	w
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

Pearson's chi-square test

- การใช้ t-test มีข้อเสียคือต้องตั้งสมมุติฐานว่าความน่าจะเป็นของกลุ่มตัวอย่างต้องกระจายตัวแบบปกติ ซึ่งในบางกรณีไม่เป็นอย่างนั้น
- วิธีการทดสอบ H_0 แบบหนึ่งที่ไม่ใช้สมมุติการกระจายตัวแบบปกติคือ χ^2 test แต่จะใช้การเปรียบเทียบความถี่ของเหตุการณ์ที่สังเกตกับความถี่ของเหตุการณ์ที่คาดว่าจะเกิดขึ้นแบบไม่ขึ้นต่อกัน

	w1 = new	w1 != new
w2 = companies	8 (new companies)	4667 (e.g., old companies)
w2 != companies	15820 (e.g., new machines)	14287173 (e.g., old machines)

C(companies) = 4675 C(new) = 15828

Pearson's chi-square test

- ปกติเราจะหาค่า X^2 แทน ซึ่งค่านี้จะ asymptotically กับ χ^2

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

- การประมาณความถี่ E_{ij} (เหตุการณ์ที่คาดว่าจะเกิดขึ้น) สามารถใช้ marginal probabilities ทำได้ เช่น E_{11} คือ

$$\begin{aligned} E_{11} &= P(new?) \times P(?companies) \times N \\ &= \frac{8+4667}{N} \times \frac{8+15820}{N} \times N \approx 5.2 \end{aligned}$$

Pearson's chi-square test

- การหา X^2 สำหรับตาราง 2x2 สามารถสรุปเป็นสมการได้ดังนี้

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

- จากสมการหาค่า X^2 สำหรับตารางในหน้าที่ 29 ได้เท่ากับ 1.55 ซึ่งตามตาราง χ^2 ที่ระดับความสำคัญ 0.05 ค่า χ^2 จะมีค่าเท่ากับ 3.841 (สำหรับ χ^2 ถ้าค่าน้อยได้ค่าระดับความสำคัญสูง)
- ดังนั้นที่ค่า X^2 เท่ากับ 1.55 จึงไม่สามารถปฏิเสธ H_0 ได้

Pearson's chi-square test

	p	0.99	0.95	0.10	**0.05**	0.01	0.005	0.001
d.f.	1	0.00016	0.0039	2.71	**3.84**	6.63	7.88	10.83
	2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
	3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
	4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
	100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

หมายเหตุ:

d.f. คำนวนได้จาก $(c - 1)(r - 1)$ โดยที่ c และ r คือจำนวนของ column และ row

Pearson's chi-square test

- สามารถใช้การจับคู่คำระหว่างคลังประโยคขนานสองภาษา (parallel corpus)

	cow	¬ cow
vache	59	6
¬ vache	8	570934

- ข้อควรระวัง: การใช้ χ^2 เหมาะสำหรับในกรณีที่มีตัวอย่างมากๆ (จำนวนตัวอย่างควรเกิน 20 ตัวอย่าง และ ความถี่ในแต่ละช่องในตารางควรเกิน 5)

Likelihood ratios

- Likelihood vs. Probability
  - Probability : $P(X \mid \Theta)$
    - ความน่าจะเป็นของเหตุการณ์ที่สนใจ เมื่อให้ เซตของตัวแปรมา
    - เช่น ถ้าทอยเหรียญที่สมดุลร้อยครั้ง ความน่าจะเป็น (probability) ที่เหรียญจะออกหัวทุกครั้งเป็นเท่าไร
  - Likelihood : $L(\Theta \mid X)$
    - ความน่าจะเป็นของตัวแปร เมื่อให้ เหตุการณ์ที่สนใจมา
    - เช่น ทอยเหรียญร้อยครั้งแล้วออกหัวทุกครั้ง ความน่าจะเป็น (likelihood) ที่เหรียญที่ทอยจะเป็นเหรียญสมดุลเป็นเท่าไร

Likelihood ratios

- $L(\Theta \mid X) \in \{\alpha\, P(X \mid \Theta) : \alpha > 0\}$
  - X คงที่
  - Θ เปลี่ยนแปลงได้

$$\frac{L(\theta_2 \mid X)}{L(\theta_1 \mid X)} = \frac{\alpha P(X \mid \theta_2)}{\alpha P(X \mid \theta_1)} = \frac{P(X \mid \theta_2)}{P(X \mid \theta_1)}$$

Likelihood ratios

- เป็นวิธีในการทดสอบสมมุติฐานที่ดีกว่า χ^2
  - สามารถใช้ในกรณี sparse data
  - ตีความผลลัพธ์ได้ง่ายกว่า
- พิจารณาการเกิดของ bigram w_1w_2 จะได้สมมุติฐาน 2 อันดังนี้
  - $H_1 : P(w_2 \mid w_1) = P(w_2 \mid \neg w_1) = p$
  - $H_2 : P(w_2 \mid w_1) = p_1, P(w_2 \mid \neg w1) = p_2, p_1 \neq p_2$
    - ถ้า H_2 เป็นจริง สมมุติให้ $p_1 >> p_2$

Likelihood ratios

- ใช้ **maximum likelihood** ในการประมาณค่า p, p_1, p_2 จาก c_1, c_2, c_{12} ซึ่งคือการนับการเกิดของ w_1, w_2, w_1w_2 ในคลังประโยค ตามลำดับ

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

- ใช้ **binomial distribution** ในการประมาณการกระจายตัว

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Likelihood ratios

	H_1	H_2
$P(w_2 \mid w_1)$	$p=c_2/N$	$p_1 = c_{12}/c_1$
$P(w_2 \mid \neg w_1)$	$p=c_2/N$	$p_2 = (c_1 - c_{12})/(N - c_1)$
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 $\neg w_1w_2$	$b(c_2 - c_{12}; N - c_1,p)$	$b(c_2 - c_{12}; N - c_1,p_2)$

Likelihood ratios

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

โดยที่ $L(k,n,x)=x^k(1-x)^{n-k}$

Likelihood ratios

$-2\log\lambda$	$C(w_1)$	$C(w_2)$	$C(w_1w_2)$	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

Likelihood ratios

$-2\log\lambda$	$C(w_1)$	$C(w_2)$	$C(w_1w_2)$	w_1	w_2
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

Likelihood ratios

- เราใช้ค่า $-2\log\lambda$ เพราะว่าค่าเป็น asymptotically กับ χ^2 นั้น หมายความว่า เราสามารถใช้ตาราง χ^2 ในการทดสอบ H_0 ได้เหมือนกับ X^2
- การอ่านผลลัพธ์ เช่น $-2\log\lambda$ ของ powerful computers มีค่าเป็น 82.96 หมายความว่า powerful computers เกิดคู่กันมากกว่าเกิดแบบสุ่ม เป็น $e^{0.5 \times 82.96} \approx 13 \times 10^{18}$ เท่า
- แม้ว่า powerful cudgels เกิดเพียง 2 ครั้งในคลังประโยค วิธีการนี้สามารถตรวจสอบได้ว่าคู่นี้อาจจะเป็น collocation

Mutual Information

- pointwise mutual information สามารถใช้ในการหา collocation ได้

$$
\begin{aligned}
I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\
&= \log_2 \frac{P(x' \mid y')}{P(x')} \\
&= \log_2 \frac{P(y' \mid x')}{P(y')}
\end{aligned}
$$

- ในทฤษฎีข้อมูล mutual information จะถูกกำหนดบนตัวแปรสุ่ม ไม่ใช่บนค่าของตัวแปรสุ่ม

Mutual Information

$I(w_1, w_2)$	$C(w_1)$	$C(w_2)$	$C(w_1w_2)$	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

Mutual Information

- ตัวอย่างการคำนวน:

$$I(\text{Ayatollah}, \text{Ruhollah}) = \log_2 \frac{\frac{20}{14307668}}{\frac{42}{14307668} \times \frac{20}{14307668}} \approx 18.38$$

- จริงๆ แล้วค่า MI ใช้ในการบอกว่า ถ้าทราบค่าของตัวแปรหนึ่งแล้วจะทราบถึงอีกตัวแปรหนึ่งได้มากขึ้นแค่ไหน
  - ไม่ได้บอกถึงการขึ้นต่อกันของเหตุการณ์ ซึ่งในหลายกรณี จะไม่เหมาะสำหรับการหาว่าสองเหตุการณ์ที่ความเกี่ยวพันและสอดคล้องกันหรือไม่

Mutual Information

- พิจารณาสองเหตุการณ์สุดโต่งนี้
  - ตัวแปรสองตัวขึ้นต่อกันเสมอ

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

  - ตัวแปรสองตัวไม่ขึ้นต่อกันเลย

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

- MI ดีในการวัดการไม่ขึ้นต่อกัน แต่ไม่ดีในการวัดการขึ้นต่อกัน

The Notion of Collocation

- Non-compositionality
  - ความหมายไม่ได้ตรงไปตรงมาตามความหมายของกลุ่มคำ อาจจะเดาไม่ได้เลย (kick the bucket) หรือ อาจจะเหลือบางส่วน (white wine)
- Non-substitutability
  - ไม่สามารถถูกแทนด้วยคำอื่นได้ เช่น เราจะไม่พูดว่า yellow wine
- Non-modifiability
  - collocation หลายตัวไม่สามารถถูกขยายได้ โดยเฉพาะ idiom เช่น เราจะไม่พูดว่า kick the old bucket

The Notion of Collocation

- เราอาจใช้วิธีการแปลคำต่อคำจากภาษาหนึ่งไปอีกภาษาหนึ่ง ในการทดสอบ collocation ได้ เช่น
  - make a decision -> สร้าง การตัดสินใจ
  - kick the bucket -> แตะ ถัง

The Notion of Collocation

- Subclass of collocations
  - Light verbs
    - make, take, do : make a decision, do a favor
  - Phrasal verbs
    - pick up, take off, turn on, etc.
  - Proper names
    - New York
  - Terminological expressions
    - hydraulic oil filter