Foundations of Statistical Natural Language Processing

5 Collocations

宮澤　彬

総合研究大学院大学 博士前期 1 年

miyazawa-a@nii.ac.jp

August 29, 2015

コロケーションとは

“Collocations of a given word are statements of the habitual or customary places of that word.” by John Rupert Firth

「ある語のコロケーションとは（個人の）習慣的あるいは（文化の）慣例的なその語の（置かれる）場所である.」

コロケーションとは

母語話者でないとなかなか分からない微妙な使い分け

- 「ご飯を炊く」と言うが「ほうれん草を炊く」とは言わない
- 「抹茶を点てる」や「風呂を点てる」などとは言うが「インスタントコーヒーを点てる」とは言わない

コロケーションは構成的でない. 表現が構成的であるとは部分の意味から全体の意味を推測できること.
不正確な表現ではあるが $[\![\cdot]\!]$ を『意味』とすれば

$$[\![\text{sunny and warm}]\!] = [\![\text{sunny}]\!] \wedge [\![\text{warm}]\!]$$

のようなものは構成的で

$$[\![\text{腹を立てる}]\!] \neq [\![\text{立てる}]\!]([\![\text{腹}]\!])$$

のようなものは構成的でない.

5.1 Frequency

コロケーションを見つけるためにはコーパス中の N グラムで発生頻度が高いものを集めればよい．しかし N グラムの頻度を単純に数えると上位がほとんど of the, in the のような機能語の組み合わせで占められてしまい，有用な情報が得られない．

これを解決する 1 つの方法は，特定の品詞の組み合わせのみを抽出することである．

Tag Pattern	Example
A N	commutative ring
N N	Banach space
A A N	stochastic differential equation
A N N	normed vector space
N A N	Jordan measurable set
N N N	probability density function
N P N	convergence in probability

New York や United States のような複合語が抽出されやすい．

5.1 Frequency

バイグラムの最初の単語を strong に限定したものと powerful に限定したものとを比較することでこれら２つの語の使い分けについて知ることができる.

w	$C(\text{strong}, w)$	w	$C(\text{powerful}, w)$
support	50	force	13
safty	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7

英作文で使いたい動詞と一緒に使うべき前置詞が分からないとき，いくつかの適当な前置詞と組み合わせたものを Google で検索して，そのヒット件数で正解を決めることと似ている.

5.2 Mean and Variance

語の共起は連続しているとは限らず，N グラムの頻度をただ数えるだけでは不十分である.

(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**

しかし，規則性がないわけではない．実際このような文脈で knocked の代わりに hit や beat は普通使わない.

5.2 Mean and Variance

離れた位置にあるコロケーションを扱うために，各単語の前後の N 個（通常 $N = 3$ または $N = 4$）の単語で組を作る．こうして得られたすべてのバイグラムを通常バイグラムと同様に扱えばよい．

Sentence: *Stocks crash as rescue plan teeters*

stocks crash	*stocks as*	*stocks rescue*		
	crash as	*crash rescue*	*crash plan*	
		as rescue	*as plan*	*as teeters*
			rescue plan	*plan teeters*

5.2 Mean and Variance

knocked と door の関係を知る方法として，２語の符号付き距離の平均と分散を求める方法がある．前頁の例文 (a)-(d) で knocked から door までの符号付き距離（door が knocked の後ろにあるときを正，前にあるときを負とする）の標本平均を計算すると

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

である．また標本分散は

$$\begin{aligned} s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(d_i - \overline{d}\right)^2 \\ &= \frac{1}{3}\left((3-4)^2 + (3-4)^2 + (5-4)^2 + (5-4)^2\right) = \frac{4}{3} \end{aligned}$$

であるが，コロケーションに関する評価では慣例的に標準偏差を用いるので，これを計算して次を得る．

$$s = \frac{2}{\sqrt{3}} \approx 1.15.$$

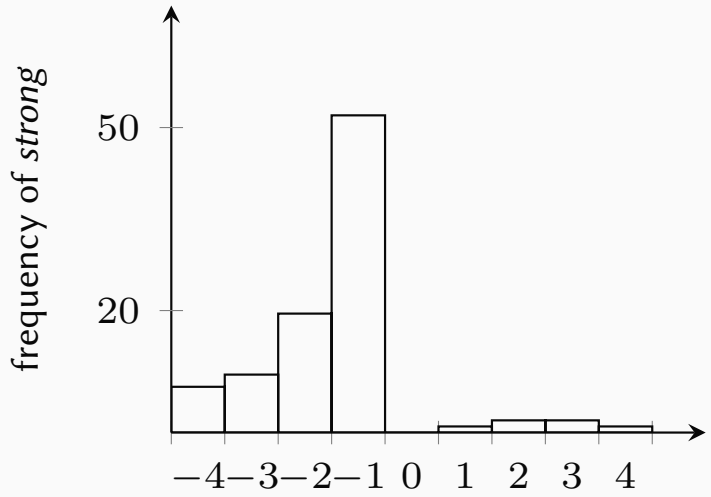
5.2 Mean and Variance

こうして得られた平均と標準偏差をどのように解釈するか？

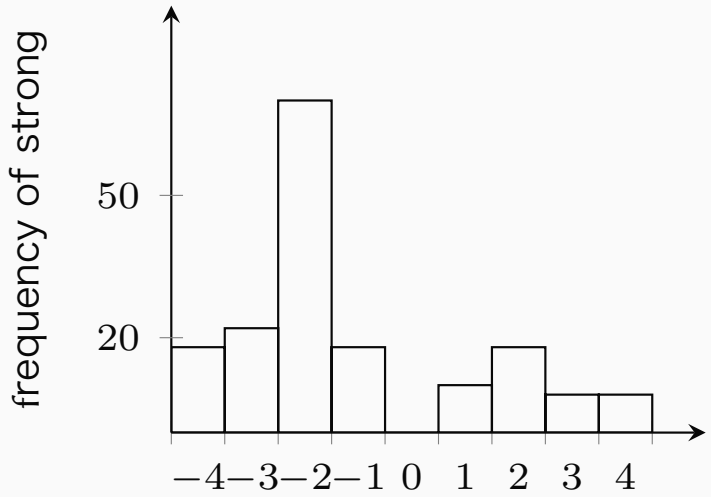
- 平均が 1 に近く，標準偏差が 0 に近い
  - 連続した 2 語として現れやすい
    New, York $\left(\overline{d} = 0.97,\ s = 0.43\right)$
- 平均が 1 よりもずっと大きく，標準偏差が 0 に近い
  - 間隔を空けて使われる表現が定型化している
    previous, games $\left(\overline{d} = 1.83,\ s = 0.48\right)$ $\rightarrow$ privious N games
- 平均が 0 に近く，標準偏差が大きい
  - あまり関連がない
    editorial, Atlanta $\left(\overline{d} = 0.44,\ s = 4.03\right)$

5.2 Mean and Variance

度数分布図を描くと直感的に分かりやすい.



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



Position of strong with respect to support $\left(\overline{d} = -1.12,\ s = 2.15\right)$

5.3 Hypothesis Testing

頻度が多い単語 2 語の組み合わせを多く見つけた場合，他の組み合わせと比較して多いかどうかが知りたい．そこでコロケーションについて検定を行うことにする.

検定のため簡単なモデルを考える．トークン 1 つを 1 回の試行とし，特定の語 w の出現を成功と捉える Bernoulli 試行とみなす．N 個のトークン（つまり N 回の試行）のうちに w が n 回出現するとする. Bernoulli 分布のパラメータ p の最尤推定量 $P(w)$ は

$$\arg\max_{\vartheta\in(0,1)} \left(\log \vartheta^n (1-\vartheta)^{N-n}\right) = \frac{n}{N}$$

で求められるから，例えば p. 166 にデータが示されている unsalted と butter の出現する確率は

$$P(\text{unsalted}) = \frac{20}{14307668}\,, \quad P(\text{butter}) = \frac{320}{14307668}$$

であると考えられる.

5.3 Hypothesis Testing

興味があるのはこれらが共起する場合である．そこで N 個のトークンの列を N 個のバイグラムの並び（文頭または文末のどちらか一方だけに特殊な記号を追加？）と捉えて，各バイグラムを 1 回の試行とし，特定のバイグラムの出現を成功とする Bernoulli 試行とみなす．これらの出現が独立であるという帰無仮説を立てて検定を行う．

$$H_0 \ : \ P(\text{unsalted butter}) = P(\text{unsalted})P(\text{butter})$$

H_0 が正しいならば，i 回目の結果を表す確率変数を $X_i : \Omega \to \{0, 1\}$ としたとき

$$X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}\left(p := P(\text{unsalted})P(\text{butter})\right)$$

である．N が大きいので de Moivre Laplace の定理により

$$Z := \frac{\sum_{i=1}^{N} X_i - Np}{\sqrt{Np(1-p)}} \overset{\text{approx.}}{\sim} \mathcal{N}(0, 1)$$

が成り立つから，以下では $Z \sim \mathcal{N}(0, 1)$ として検定を行う．

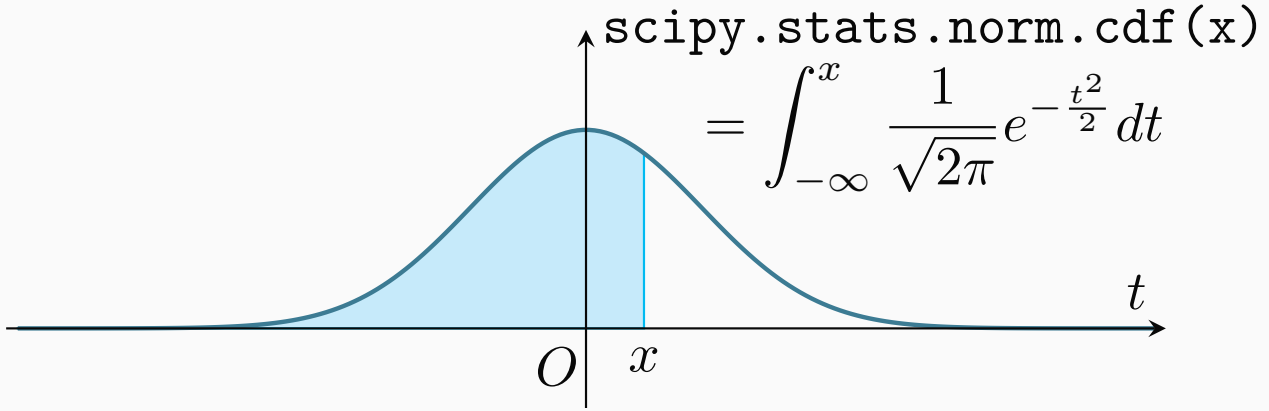
5.3.1 The t test

p が 0 に近いので $p(1-p) \simeq p$ が成り立つ．ゆえに観測値を $x_1, \ldots, x_n$ とすると

$$Z \simeq \frac{\sum_{i=1}^{N} x_i - Np}{\sqrt{Np}} = \frac{20 - 24 \cdot 320/14307668}{\sqrt{24 \cdot 320/14307668}} \approx 863$$

となる．以下の計算により H_0 は有意水準 1%で棄却される．したがって unsalted butter はコロケーションであるといえる．

```
>>> import scipy.stats
>>> scipy.stats.norm.cdf(863)
1.0
```



5.3.1 The t test

教科書では t 値を用いて検定を行なっているが,

- $X_1, \ldots, X_N$ がそれぞれ独立に同一の正規分布に従う.
- 母分散 σ^2 が未知である.

が成立していないので t 値を使うのは不適当である.

しかし

> When the expected cooccurrence frequency E_{11} (under H_0) is small, z-score values can become very large, leading to highly inflated scores for low-frequency pair types. The cause of this problem can be traced back to the denominator of the z-score equation, where E_{11} is the (approximate) variance of X_{11} under the point null hypothesis of independence.
>
> http://www.collocations.de/AM/section4.html

5.3.1 The t test

さらに

> The t test and other statistical tests are most useful as a method for ranking collocations. The level of significance itself is less useful. In fact, in most publications that we cite in this chapter, the level of significance is never looked at. All that is used is the scores and the resulting ranking.
>
> Foundations of Statistical Natural Language Processing p. 166

5.3.2 Hypothesis testing of differences

母集団の平均の差に関する検定（量）により，似たような意味を持つ２つの語の使い分けについて知ることができる．$\mathcal{N}(\mu_x, \sigma_x^2)$ に従う母集団からの大きさ N_x の標本平均を $\overline{X}$，$\mathcal{N}(\mu_y, \sigma_y^2)$ に従う母集団からの大きさ N_y の標本平均を $\overline{Y}$ とする．正規分布の重ね合わせの性質より，

$$Z := \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma_x^2}{N_x} + \dfrac{\sigma_y^2}{N_y}}} \sim \mathcal{N}(0,\, 1)$$

が成り立つ．帰無仮説を $H_0 : \mu_x = \mu_y$ として検定を行いたいところであるが，前頁で見たように検定にはあまり興味がない．知りたいのは $\mu_x = \mu_y$ のときの Z である．今回の設定では，比較したい（似ている）単語を v^1, v^2 とし，それ以外の各単語 w について

$$Z \simeq \frac{P\left(v^1 w\right) - P\left(v^2 w\right)}{\sqrt{\dfrac{P\left(v^1 w\right)}{N_x} + \dfrac{P\left(v^2 w\right)}{N_y}}} = \frac{C\left(v^1 w\right) - C\left(v^2 w\right)}{\sqrt{C\left(v^1 w\right) + C\left(v^2 w\right)}}$$

を計算する．ここでは $N_x = N_y$ であることを使った．

5.3.2 Hypothesis testing of differences

Z	$C(w)$	$C(\text{strong } w)$	$C(\text{powerful } w)$	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
		⋮		
4.0249	1093	19	1	opposition
4.5825	3741	21	0	sales
4.6904	986	22	0	safety
6.3257	3616	58	7	enough
7.0710	3685	50	0	support

このランキングから powerful が computer のような実体のあるものの強さを表し，strong が support のような内面的な強さを表している傾向をつかむことができる.

5.3.3 Pearson's chi-square test

t 検定が使えるのは母集団が正規分布していると仮定できる場合のみである．しかし実際のところ，普通はそのように仮定することができない．代替手段として分割表を用いた独立性の検定をコロケーションかどうかの判定に使うことを考える．

	X	$\neg X$	
Y	a	b	$a+b$
$\neg Y$	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

5.3.3 Pearson's chi-square test

$$\chi^2 := \frac{\left(a - \frac{(a+b)(a+c)}{a+b+c+d}\right)^2}{\frac{(a+b)(a+c)}{a+b+c+d}} + \frac{\left(b - \frac{(a+b)(b+d)}{a+b+c+d}\right)^2}{\frac{(a+b)(b+d)}{a+b+c+d}}$$
$$+ \frac{\left(c - \frac{(c+d)(a+c)}{a+b+c+d}\right)^2}{\frac{(c+d)(a+c)}{a+b+c+d}} + \frac{\left(d - \frac{(c+d)(b+d)}{a+b+c+d}\right)^2}{\frac{(b+d)(a+c)}{a+b+c+d}}$$
$$= \frac{\frac{(ad-bc)^2}{(a+b+c+d)^2}}{\frac{(a+b)(a+c)}{a+b+c+d}} + \frac{\frac{(ad-bc)^2}{(a+b+c+d)^2}}{\frac{(a+b)(b+d)}{a+b+c+d}}$$
$$+ \frac{\frac{(ad-bc)^2}{(a+b+c+d)^2}}{\frac{(c+d)(a+c)}{a+b+c+d}} + \frac{\frac{(ad-bc)^2}{(a+b+c+d)^2}}{\frac{(b+d)(a+c)}{a+b+c+d}}$$

5.3.3 Pearson's chi-square test

$$\begin{aligned}\chi^2 &= \frac{1}{(a+b+c+d)(a+b)(c+d)(a+c)(b+d)} \\ &\quad ((ad-bc)^2((c+d)(b+d)+(c+d)(a+c)+ \\ &\quad (a+b)(b+d)+(a+b)(c+d))) \\ &= \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}\end{aligned}$$

5.3.3 Pearson's chi-square test

帰無仮説として

$$H_0 : \text{unsalted と butter の出現は独立である.}$$

を設定する.

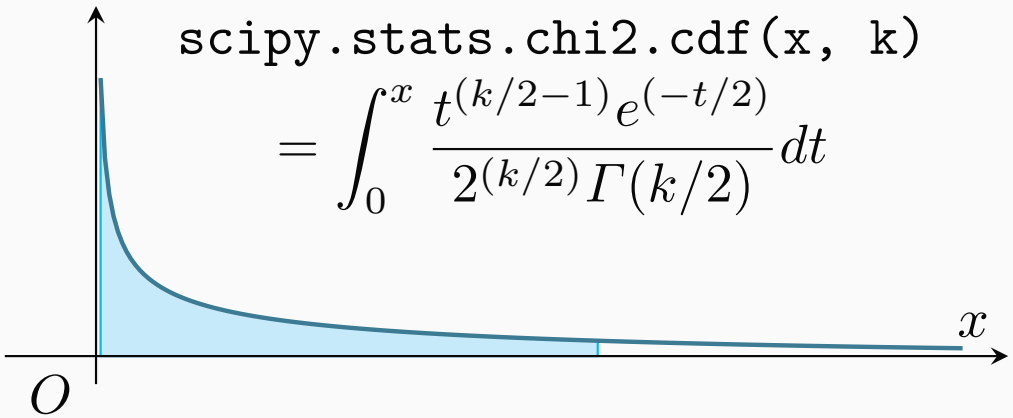
	unsalted	¬unsalted	
butter	20	300	320
¬butter	4	14307344	14307348
	24	14307644	14307668

$$\chi^2 = \frac{14307668(20 \cdot 14307344 - 300 \cdot 4)^2}{320 \cdot 14307348 \cdot 14307644 \cdot 24} \approx 745169$$

χ^2 が $\mathrm{df} = (2-1)(2-1) = 1$ の χ^2 分布における片側 1%点 6.63 よりも大きいので，H_0 は棄却される．つまり unsalted と butter の出現は独立とはいえず，したがって unsalted butter はコロケーションであるといえる.

5.3.3 Pearson's chi-square test

```
>>> import numpy as np
>>> import scipy.stats
>>> chi2, p, dof, expected = \
        scipy.stats.chi2_contingency(np.array(
        [[20,300],[4,14307344]]),correction=False)
>>> chi2, p, dof
(745168.95831120119, 0.0, 1)
```



カイ二乗検定（量）のコロケーション以外の応用として，コーパスの類似度の計測や，対訳コーパスの中の訳語の発見などがある．

5.3.4 Likelihood ratios

疎なデータに対してはカイ二乗検定より，尤度比を使ったほうがよい (Dunning 1993).

	w_2	$\neg w_2$	
w_1	c_{12}	$c_1 - c_{12}$	c_1
$\neg w_1$	$c_2 - c_{12}$	$(N - c_1) - (c_2 - c_{12})$	$N - c_1$
	c_2	c_2	N

$p_1 := P(w_2|w_1)$, $p_2 := P(w_2|\neg w_1)$ とするとき,

$$H_1 : p_1 = p_2 \,, \quad H_2 : p_1 \neq p_2$$

という仮説を立てる.

5.3.4 Likelihood ratios

H_2 が正しいと仮定するとすると最尤推定量は以下のようになる.

$$
\begin{aligned}
L(H_2) := & \binom{c_1}{c_{12}} p_1^{c_{12}} (1-p_1)^{c_1 - c_{12}} \\
& \binom{N-c_1}{c_2 - c_{12}} p_2^{c_2 - c_{12}} (1-p_2)^{(N-c_1)-(c_2-c_{12})} \\
= & \binom{c_1}{c_{12}} \left(\frac{c_{12}}{c_1}\right)^{c_{12}} \left(1 - \frac{c_{12}}{c_1}\right)^{c_1 - c_{12}} \\
& \binom{N-c_1}{c_2 - c_{12}} \left(\frac{c_2 - c_{12}}{N-c_1}\right)^{c_2 - c_{12}} \\
& \left(1 - \frac{c_2 - c_{12}}{N-c_1}\right)^{(N-c_1)-(c_2-c_{12})}.
\end{aligned}
$$

5.3.4 Likelihood ratios

一方 H_1 が正しいと仮定したとき最尤推定量は

$$
\begin{aligned}
L(H_1) := & \binom{c_1}{c_{12}} p^{c_{12}} (1-p)^{c_1 - c_{12}} \\
& \binom{N-c_1}{c_2-c_{12}} p^{c_2-c_{12}} (1-p)^{(N-c_1)-(c_2-c_{12})} \\
= & \binom{c_1}{c_{12}} \left(\frac{c_2}{N}\right)^{c_{12}} \left(1-\frac{c_2}{N}\right)^{c_1-c_{12}} \\
& \binom{N-c_1}{c_2-c_{12}} \left(\frac{c_2}{N}\right)^{c_2-c_{12}} \left(1-\frac{c_2}{N}\right)^{(N-c_1)-(c_2-c_{12})}
\end{aligned}
$$

である.

H_1 と H_2 の尤度比を $\lambda := L(H_1)/L(H_2)$ とすると，$-2\log\lambda$ は漸近的に自由度 1 のカイ二乗分布に従うことが知られている．$-2\log\lambda$ は H_2 が尤もらしいとき大きい値をとる.

5.3.4 Likelihood ratios

$-2\log\lambda$	$C(w^1)$	$C(w^2)$	$C(w^1w^1)$	w^1	w^2
−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	

尤度比と似たような概念に相対頻度比がある．２つのコーパスの N グラムにおいて相対頻度比が 1 から大きく離れたものは、主題を特徴付ける N グラムであると考えられる.

Ratio	1990	1989	w^1	w^2
0.0372	2	68	East	Berliners
0.0482	2	34	EAST	GERMANS

5.4 Mutual Information

情報理論において相互情報量 Mutual Infomation (MI) は確率変数（の分布）に対して定義される．しかしここでは特定の事象についての情報量 Pointwise Mutual Information (PMI) を用いる．

$$I(x, y) := \log_2 \frac{P(xy)}{P(x)P(y)} \tag{5.11}$$

$$= \log_2 \frac{P(x|y)}{P(x)} \tag{5.12}$$

$$= \log_2 \frac{P(y|x)}{P(y)} \tag{5.13}$$

5.4 Mutual Information

事象 x の情報量 $I(x)$ は

$$I(x) := \log \frac{1}{P(x)}$$

で定義されるのだった．これは x が稀な事象であれば大きな値をとる．$I(x, y)$ を (5.12) に従って解釈すると，事象 y が起こると教えられれば，x が起こることについて持っている情報量が $I(x, y)$ だけ増えるということである．

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

5.4 Mutual Information

PMI は対応関係の指標として使えるのか？（その１）

フランス語において

- chambre /ʃɑ̃bʁ/ 部屋，議会

あることに注意すれば，英語・フランス語の aligned corpus において英文中で house という語が出てくると知れば，仏文中で chambre が出てきそうだと予想がつく．よって $I(\text{chambre}|\text{house})$ は大きいと考えられる．しかし Hansard corpus（カナダの議会の議事録から作られたコーパス）において

$$P(\text{house}|\text{chambre}) < P(\text{house}|\text{communes})$$

すなわち

$$I(\text{chambre},\ \text{house}) < I(\text{communes},\ \text{house})$$

となってしまう．こうなってしまうのは Hansard corpus において house の最もよく使われる用法が House of Commons（英国の下院）であり，それに対応するフランス語が chambre des communes だからである．

5.4 Mutual Information

PMI は対応関係の指標として使えるのか？（その２）

1. x と y が独立なとき

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

   よさそう.
2. x の出現が y の出現と一致するとき

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

   $I(x, y)$ は極めて頻度の低い事象で構成されているときに大きい値になってしまう．ある程度の出現する単語でないと判断材料が足りない.

あまりにも低い頻度の単語については除外する方法が考えられるが，頻度の低い語で構成されているに大きな値をとる根本的な問題が解決できていない．$C(w^1w^2)I(w^1w^2)$ のように補正をかける方法などが提案されている.

5.5 The Notion of Collocation

コロケーションの定義をどう定めるか.

- Non-compositionarity
  - white wine が「白いワイン」と言われれば違う気がする.
- Non-substitutability
  - ?「腹」と「お腹(なか)」は同じ意味だと考えられるが,「お腹を立てる」とは言わない.
- Non-modifiability
  - 「口を開く」(話し始める)を「口を大きく開く」などとは言わない.

doctor - nurse のように 関連する語の組も含めることがある. 固有名詞の組合わせ(姓名など)は, 語彙的なコロケーションではないが, NLP の文脈では含めることが多い.