Log-r スコアの提案に基づく英語 Bigram の分析

藤村逸子（名古屋大学）・青木繁伸（群馬大学）

Characterization of English Bigrams Based on Log-r

Itsuko FUJIMURA (Nagoya University), Shigenobu AOKI (Gunma University)

Abstract

A new score called Log-r is proposed, which is more effective than the commonly used Mutual Information (MI) for measuring the degree of non-compositionality of bigrams. This conclusion is drawn from the evaluation of one million bigrams taken from a huge English corpus of 1.1 billion words. While the Log-r represents only the degree of non-compositionality of a bigram, the MI measures the combination of the degree of non-compositionality and the frequency. A three-dimensional analysis of each bigram can be made with its Log-r, its raw frequency, and the approximate order in the corpus of its constituent words. This last is calculated according to Zipf's law. Robust and transparent analysis of collocations could be conducted using this procedure.

1. はじめに

大規模コーパスに基づく言語研究のひとつとしてコロケーションの研究が盛んにおこなわれている。コロケーションは語と語の慣用的な結びつきと定義されるが、それには種々のタイプのものが含まれる。それぞれのタイプを特徴づける基本的な特性として言及されることが多いのは、連語[1]の粗頻度と、連語を構成する単語間の結びつきの強度の 2 つである（Ellis, 2012; Gries, 2012; Wray, 2012）。粗頻度はわかりやすい特性であるが、結びつきの強度は degree of association、degree of non-compositionality、degree of fixedness、degree of coherence などと呼ばれ統一的に扱われてはいない。また、その計測法としては MI スコア（Church & Hanks, 1990）に言及されることが多いが、一方で種々の他の計測法が提案されるなど（Pecina, 2008; 相澤・内山, 2011）研究はいまだ途上にある。本研究では、2 語連語の結びつきの強度をはかる簡素な指標として Log-r スコアを提案し、MI と対照させつつ、言語現象としてのコロケーションを理解する上のその有用性を主張する。以下ではまず、Log-r の数学的根拠を説明する。次に、英語の 2 語連語を材料にして散布図を描き、MI と Log-r との相違を表示する。また Log-r、2 語連語の粗頻度、2 語連語を構成する単語の親密度の 3 つによって、英語の 2 語連語のいくつかを特徴づける。

2. Log−r スコアの数学的根拠

2 語の結びつきの強さを示す指標として、2 変数（単語 x と単語 y）の属性相関を表す相関係数rの常用対数を提案し、それを Log-r と名づける[2]。ピアソンの積率相関係数rの定義式は

---

1 ここで連語とは、その共起の慣用性に関わらず単に語の連続を指す。2 語連語（bigram）は 2 語の連続を指す。

2 2 語の結びつきを測る指標としてすでに提案されている、z スコア、カイ二乗値（χ^2）、phi 係数は、rと共通し

以下のとおりである。

$$r = \frac{x,y \text{ の共分散}}{x \text{ の標準偏差} \times y \text{ の標準偏差}}$$

本研究では、ポワソン分布を仮定して次の近似式を使う。なおf_{xy}は連語 xy の頻度、f_xとf_yは単語 x、単語 y の頻度である。

$$r \fallingdotseq \frac{f_{xy}}{\sqrt{f_x f_y}}$$

対数をとる理由は、語の頻度が Zipf の法則に従う極端に範囲の広い統計量だからである（概算では、単語のコーパス内での出現率（%）＝10/順位）。対数をとって比較すると数値の処理が容易になり、言語現象を把握しやすくなる。

MI スコアは、単語 x と単語 y が共起する確率を偶然の共起の確率と比較したものとされ、2 語連語の結びつきの強度の指標として言及されることが多い（Ellis, 2012; Gries, 2012; Wray, 2012）。また、粗頻度と並ぶ基本的な尺度とも言われている（相澤・内山, 2011[3]）。

$$MI = \log_2 \frac{f_{xy}\, N}{f_x\, f_y}$$

しかし MI は単純に x と y の結びつきの強度を示すのではない。総語数Nが同じで、f_x: f_y: f_{xy}の割合が同じでも、xy の頻度によって MI の値はかわる。xy の頻度が高いほど MI の値は低く、xy の頻度が低いほど MI は高い[4]。たとえば英語の「色彩語＋名詞」8000 種類（大文字・小文字を区別）の中で、語頭が大文字の White House の MI は 507 位であり、red bandana よりも後にある。White House は red bandana に比べて、頻度が圧倒的に高いために MI の値は低くなっている。一方、Log-r では White House は 1 位である。直観的には White House（ホワイトハウス）の 2 語間の結びつきは red bandana（赤いバンダナ）のそれより強く、結びつきの強度を示す指標として Log-r は MI より適切と思われる。

3. Log-r と MI による散布図の比較

以下では、大規模コーパスから得た 2 語連語の散布図を描くことにより、Log-r と MI の特徴を明らかにする。コーパスは総語数約 11 億語の英語新聞であり、頻度 54 回以上の 2 語連語 1,039,996 種類をデータとする。これらの 2 語連語のそれぞれに対して、粗頻度の常用対数 Log(f_{xy})、Log-r、MI を計算し、Wray（2012, 241）の図を参考にして横軸に Log(f_{xy})をとり、縦軸に Log-r（図 1）と MI（図 2）をとった散布図を描く。Log-r は粗頻度からは独立した指標である（図 1）。散布図の下端が右上方向の斜線になっているのは、単語の数は有限のため、2 語連語の頻度が高くなるにつれて、弱い結びつきの連語は存在しなくなるという現実に対応している。一方、MI は結びつきの強さと頻度が融合した指標であり、計算式の特長のゆえに、頻度が高くなるにつれ MI の値は低くなる（図 2）。

た性質をもっている（cf. Pecina, 2008; 相澤・内山, 2011）。すなわち Log-r は全く新しい指標というわけではない。

[3] 自己相互情報量（PMI）として言及されているが、対数の底が異なるだけであり、MI と本質的に同一である。

[4] コーパスの大きさによっても MI の値はかわる。

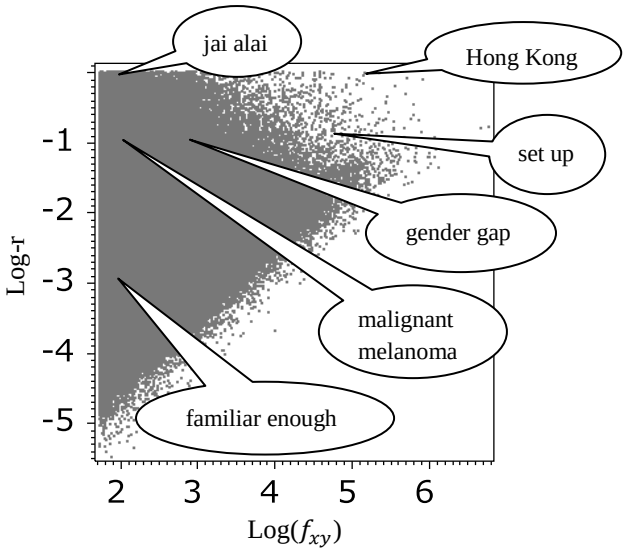

図 1　Log(f_{xy})と Log-r の散布図(110 万件)

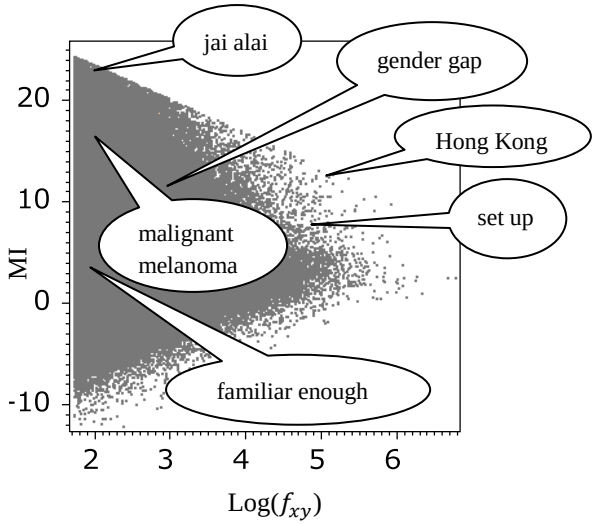

図 2　Log(f_{xy})と MI の散布図(110 万件)

図 1 と図 2 の上に、例として 6 個の 2 語連語を配置した（cf. 表 2)。2 つの図を比較すると、MI は Log-r を右下がりに歪曲させた指標であることがわかる。Log-r は相関係数の対数なので、粗頻度と同じく簡素で透明性の高い堅牢な指標であり、粗頻度と組み合わせることにより連語の特徴を明示的に示すことができる。

4. Log−r の散布図の各部分の特徴

表 1 は、総語数 10 億語のコーパスにおける、頻度の等しい x と y からなる 2 語連語 xy を仮定して、Log-r の散布図（図１）の各部分の特徴をシミュレートした結果である。「x の概算順位」は Zipf の法則により求め、その順位に基づき「x の親密度」を推定した。これにより、散布図上の 2 語連語の特徴を、2 語の結びつきの強度(=Log-r)、2 語連語の頻度度（=Log(f_{xy})）、単語の親密度（=x の概算順位）の３つの指標によって記述することができる。

5. 英語の 2 語連語の分析

以上で提案した指標により、図 1 と図 2 に表示した 6 つの英語の 2 語連語の特徴を表 2 に示す。+の数はそれぞれの尺度の程度が高いことを示す。たとえば Hong Kong は 2 語の結びつきは最強で頻度も高い。jai alai は結びつきは最強であるが、連語の頻度は極めて低く、それぞれの単語も誰も知らないような語である。set up は結びつきは強く頻度は高く、それぞれの単語は機能語のレベルである。

6. まとめ

このように、堅牢な統計量である Log-r を利用し、粗頻度、Zipf の法則などの同様に堅牢な統計量を組み合わせて連語を計測することにより、コロケーションのタイプを明示的に記述し、説明することが可能になる。MI は、頻度が低く結びつきの強い連語を採取するための実用的ツールとしての意味はあるが、コロケーションを説明するための指標として有効とは言えない。超大規模なデータを用いることによって以上が明らかとなった。

表１　Log(f_{xy})と Log-r の散布図(図１)の各部分

Log-r	Log(f_{xy})	f_{xy}	f_x (=f_y)	r	f_x/N	x の概算順位	x の親密度	MI
0	2	100	100	1	0.0000001	1000000	未知語	23.3
-1	2	100	1000	0.1	0.000001	100000	希少語	16.6
-2	2	100	10000	0.01	0.00001	10000	高級語	10.0
-3	2	100	100000	0.001	0.0001	1000	基本語	3.3
-4	2	100	1000000	0.0001	0.001	100	機能語	-3.3
-5	2	100	10000000	0.00001	0.01	10	超機能語	-10.0
0	3	1000	1000	1	0.000001	100000	希少語	19.9
-1	3	1000	10000	0.1	0.00001	10000	高級語	13.3
-2	3	1000	100000	0.01	0.0001	1000	基本語	6.6
-3	3	1000	1000000	0.001	0.001	100	機能語	0.0
-4	3	1000	10000000	0.0001	0.01	10	超機能語	-6.6
0	4	10000	10000	1	0.00001	10000	高級語	16.6
-1	4	10000	100000	0.1	0.0001	1000	基本語	10.0
-2	4	10000	1000000	0.01	0.001	100	機能語	3.3
-3	4	10000	10000000	0.001	0.01	10	超機能語	-3.3
0	5	100000	100000	1	0.0001	1000	基本語	13.3
-1	5	100000	1000000	0.1	0.001	100	機能語	6.6
-2	5	100000	10000000	0.01	0.01	10	超機能語	0.0
0	6	1000000	1000000	1	0.001	100	機能語	10.0
-1	6	1000000	10000000	0.1	0.01	10	超機能語	3.3

表２　2 語連語の特徴付け

2語連語	2語の結びつきの強度・Log-r		連語の頻度・Log(f_{xy})		単語の親密度・概算順位		参考：MI
Hong Kong	+++++	−0.01	+++++	5.16	++++	1000	12.82
jai alai	+++++	−0.02	++	2.18	+	1000000	22.63
set up	++++	−1.01	++++	4.90	+++++	100	7.04
malignant melanoma	++++	−1.02	++	2.04	++	100000	16.73
gender gap	+++	−1.20	+++	3.08	+++	10000	11.86
familiar enough	++	−2.97	++	2.05	++++	1000	3.47

参考文献

Church, Kenneth & Hanks, Patrick (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics, 16(1)*, 22-29.

Ellis, Nick. C. (2012). Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics, 32*, 17-44.

Gries, Stefan. Th. (2012). Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language, 36(3)*, 477-510.

Pecina, Pavel (2008). A Machine Learning Approach to Multiword Expression Extraction, In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Wray, Alison (2012). What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play, *Annual Review of Applied Linguistics, 32*, 231-254.

相澤彰子・内山清子 (2011).「語の共起と類似性」松本裕治(編)『言語と情報科学』朝倉書店 58-76.