平均場近似を用いたランダムなしりとりの解析

藤田悠朔 [1], 鈴木岳人 [2], 水口毅 [1]

[1] 大阪公立大学 大学院理学研究科 物理学専攻
[2] 高千穂大学 人間科学部

概要

しりとりは前の単語に、その末尾文字から始まる単語をつなげるゲームである。本研究ではしりとりで使用できる単語を定め、その中からランダムに単語を選ぶことを考える。我々は単語列の長さ（鎖長）に着目し、平均場近似を施した場合の鎖長分布に関する理論的な解析を行った。

Mean-field Analysis of Random Word Chain Games

Yusaku Fujita [1], Takehito Suzuki [2], Tsuyoshi Mizuguchi [1]

[1] Department of Physics, Graduate School of Science, Osaka Metropolitan University
[2] Faculty of Human Sciences, Takachiho University

Abstract

A word chain game is a game where players connect words beginning with the last character of the previous word. In this study, a “dictionary”, namely a set of usable words in the game, is defined, and a random word-selecting process in the dictionary is considered. We focused on the chain length of each trial and analysed its distribution with a mean-field approximation.

1 はじめに

しりとりは、“りんご”→“ごりら”→“らっぱ”→··· のように、前の単語にその末尾文字から始まる単語をつなげていくゲームである。他のルールとしては、(i) 一度使用した単語は再度使用できない (ii) 次の単語がなくなったら終了（負け）がある。本研究では、使用できる単語の集合（辞書）を定め、ゲームの勝敗や戦略を度外視したランダムなしりとりを考える。すなわち、単語は残されている選択可能な単語の中からランダムに選び、選択された単語は辞書から消す、という過程を終了するまで繰り返す。こうしてできる鎖長（しりとりの単語列の長さ）がどのようになるのかに着目した。

図 1 は文学作品 “Moby-Dick” [1] に登場する英語の名詞 19088 語で 50000 回ランダムなしりとりを実行したときの鎖長の分布である。この非対称な分布はどのように得られるのかという疑問が提起される。しかし Moby-Dick 辞書には 26 文字が使われており、分布を求めるのは容易ではない。本論文では、使用されている文字の数が 2 という単純な場合に限定し、その鎖長分布について、近似を用いた解析結果を報告する。

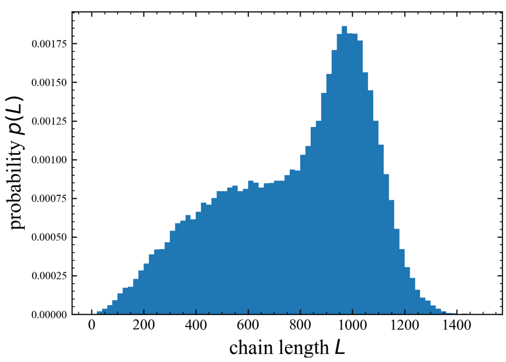

図 1: Moby-Dick 辞書の鎖長分布（$N = 50000$）

2 解析手法

辞書の単語数を D、しりとりに使われる文字の数を C とすると、辞書は文字を頂点（頂点数 C）、単語を辺（リンク数 D）とする多重有向ネットワークを構成する。そして、ランダムなしりとりはそのネットワーク上の自己回避ランダムウォーク (SAW) に対応する。図 2 に小さな辞書ネットワーク ($D=8, C=3$) 上の長さ $L=5$ の SAW の例を示す。

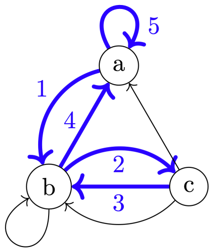

図 2: 青矢印は単語列 “adverb” → “basic” → “climb” → “banana” → “area” に対応、黒矢印は使われなかった単語を表す。

文字 θ の入次数を $k_{\mathrm{in},\theta}$、出次数を $k_{\mathrm{out},\theta}$ とすると、しりとりは次の式を満たす文字 θ で終了する。

$$k_{\mathrm{in},\theta} \geq k_{\mathrm{out},\theta}. \tag{1}$$

(1) 式を満たす文字を終了可能文字と呼び、満たさない文字を終了不能文字と呼ぶ。そして、しりとりが鎖長 L かつ文字 θ で終了する確率を $p_\theta(L)$ で表す。

以下では $p_\theta(L)$ そのものではなく、平均場近似を施した鎖長分布 $\tilde{p}_\theta(L)$ を取り扱う。平均場近似とは、与えられた辞書の各文字の入次数と出次数を保ったまま、頂点同士をランダムにつなぎかえてできる一連の辞書群（シャッフル辞書群）に対する平均を意味する。

3 結果

$C=2$ のシャッフル辞書群の鎖長分布 $\tilde{p}_\theta(L)$ は負の超幾何分布に従うことが判明した。終了可能文字を x、終了不能文字を z とすると、$\tilde{p}_x(L)$ は次のように書くことができる。

$$\begin{aligned}\tilde{p}_x(L) = &\frac{1}{2}\frac{\binom{L-1}{k_{\mathrm{out},x}-1}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}}}{\binom{D}{k_{\mathrm{in},x}}} \\ &+ \frac{1}{2}\frac{\binom{L-1}{k_{\mathrm{out},x}}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}-1}}{\binom{D}{k_{\mathrm{in},x}}}.\end{aligned} \tag{2}$$

L/D をカバーレートと呼ぶ。(2) 式と負の超幾何分布の統計量 [2] をもとに、カバーレートの最大値 $L_{\max}/D$、最小値 $L_{\min}/D$、平均値 L_{mean}/D を計算すると、

$$\frac{L_{\max}}{D} = 1 + \frac{k_{\mathrm{out},x} - k_{\mathrm{in},x} + 1}{D} \tag{3}$$

$$\frac{L_{\min}}{D} = \frac{k_{\mathrm{out},x}}{D} \tag{4}$$

$$\frac{L_{\mathrm{mean}}}{D} \simeq \frac{k_{\mathrm{out},x}}{k_{\mathrm{in},x}} \tag{5}$$

と求められる。ただし、平均値の計算には $D \gg 1$, $k_{\mathrm{in},x} \gg 1, k_{\mathrm{out},x} \gg 1$ の近似を用いた。数値計算でも (3)–(5) 式が成立することが確認された。図 3 は平均値についての計算結果の一例であり、D が大きくなるにつれて (5) 式の関係に漸近している。

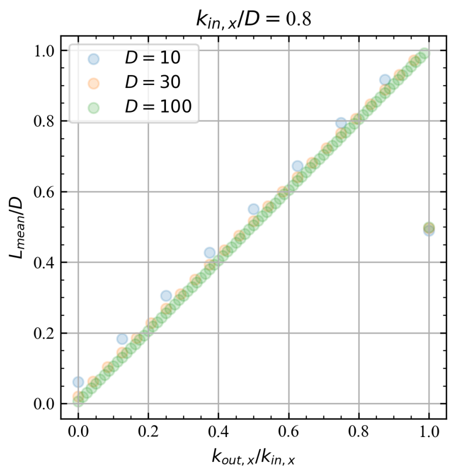

図 3: $\dfrac{L_{\mathrm{mean}}}{D}$ の $\dfrac{k_{\mathrm{out},x}}{k_{\mathrm{in},x}}$ 依存性 $\left(\dfrac{k_{\mathrm{in},x}}{D} = 0.8 \text{ に固定}\right)$

参考文献

[1] Project Gutenberg, https://www.gutenberg.org/ebooks/2701

[2] R. A. Khan, Sankhyā: The Indian Journal of Statistics, B(1960-2002), **56**(1994) 309.