複雑ネットワークの視点による二字熟語の構造について

山本健（中央大学理工学部）

Structure of Two-letter Kanji Compounds Based on Complex Network Analysis

Ken Yamamoto (Faculty of Science and Engineering, Chuo University)

1　複雑ネットワーク

複雑ネットワークは、多数の要素の「つながり方」について調べる道具として発展してきた。人と人との知人関係、インターネットのルータの接続関係、および企業間の取引関係など様々なシステムが複雑ネットワークの視点でモデル化され分析されている。複雑ネットワークを構成する要素を**頂点**、頂点どうしのつながりを**辺**とよぶ。

数珠つなぎに知り合いをたどっていくと、少ない人数を経由するだけで誰にでも行き着けるという都市伝説がある[1]。複雑ネットワークのことばでは、ある頂点から辺をたどって別の頂点に行くための最短距離（経路長）が小さいということである。現実の多くのネットワークでも平均的な経路長が小さいことが知られている。また、「自分の知人の知人」にあたる人物が自分の知人である可能性は、世界中からでたらめに選んだ人物が自分の知人である可能性よりも格段に大きい。つまり、知人関係のネットワークには「三角形」が多く含まれるということである。ネットワーク中の「三角形」の多さを測る量が**クラスター係数**である。ネットワークが三角形を全く含まなければクラスター係数は 0 となり、ネットワークのどの 3 頂点も三角形をつくっていればクラスター係数は 1 になる。平均経路長が小さくクラスター係数が大きいとき、ネットワークは**スモールワールド**であるという。

ネットワークの統計的指標として**次数分布**がある。ちょうど k 本の辺が接続している頂点（次数が k である頂点）の割合を $P(k)$ で表し、これを次数 k に関する確率分布とみなしたとき、$P(k)$ を次数分布とよぶ。次数分布がベキ乗関数

$$P(k) \propto k^{-\gamma}$$

で表されるようなネットワークを、**スケールフリー**ネットワークとよぶ。$k \to \infty$ におけるベキ乗関数の減衰は指数関数や正規分布関数よりも緩慢であり、スケールフリーネットワークでは、次数がとても大きな頂点が存在しやすいという特徴がある。

Watts, and Strogatz (1998) によるスモールワールド性および Barabási, and Albert (1999) によるスケールフリー性の発見が、近年の複雑ネットワーク研究の潮流をつくるきっかけとなった。複雑ネットワークの諸概念をまとめた成書としては、増田、今野 (2010), 矢久保 (2013), Newman (2010) など豊富に刊行されている。以下、「複雑ネットワーク」のことを単に「ネットワーク」とよぶこととする。

本稿で取り扱うのは、二字熟語を「2 つの漢字を結ぶ辺」とみなして構成されるネットワークである。このネットワークがスモールワールドかつスケールフリーであることを示し、これらの特徴を定量的に再現するシンプルなネットワーク生成モデルを提案する。

[1] いわゆる「6 次の隔たり」

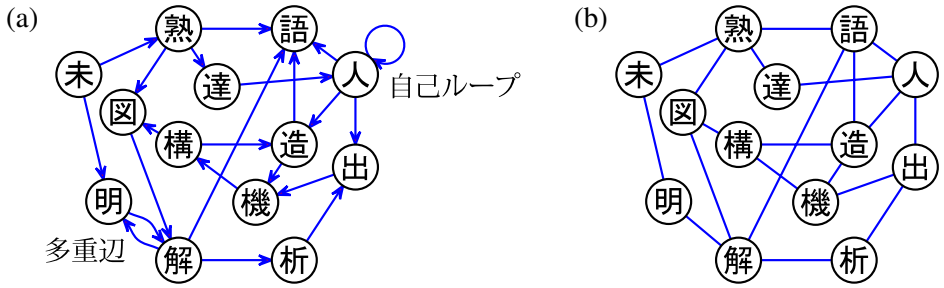

図 1: 二字熟語ネットワークの一部分。(a) 辺は向きをもち、ネットワークには自己ループおよび多重辺が存在する。(b) 本研究では辺の向きを無視して単純化したネットワークに注目する。

2　二字熟語ネットワーク

図 1 は二字熟語ネットワークの一部分である。例えば、「構造」という熟語に対応して「構」と「造」を結ぶ辺が引かれている。漢字の順序（1 字目か 2 字目か）があるので、ネットワークの辺には自然に向きがつけられる。「人人」という熟語のように、ある頂点から自分自身へ向かう辺（自己ループ）が存在する。また、「明解」と「解明」のように複数の辺をもつ頂点の対も存在する。しかし本研究では、漢字と漢字の最も単純なつながり方に関する特徴を明らかにするという点から、辺の向きは考えないこととする。自己ループを消去し、多重辺は 1 本にまとめたネットワーク（無向単純ネットワーク）について調べる（図 1 (b)）。

二字熟語のデータとして 3 冊の国語辞典『広辞苑　第四版』、『岩波国語辞典　第五版』、および『三省堂国語辞典　第 4 版』を利用した。それぞれの国語辞典の見出し語から漢字 2 文字のものをすべて抜き出し、対応するネットワークの解析をおこなった。分析結果を表 1 に示す。ネットワークの規模を特徴づける頂点数（= 漢字の種類）、辺数（≈ 熟語の個数[2]）、および平均次数（1 つの漢字が平均して何個の熟語に使用されているか）はいずれも広辞苑が他の 2 冊より大きくなっている。これは、直感的な「辞書の規模」と一致する結果である。平均経路長 l が小さく、クラスター係数 C がランダムネットワークで期待される値 C_{random} と比べてはるかに大きいことから、3 つのネットワークはスモールワールドである。これらの二字熟語ネットワークはスケールフリーでもあり、3 つの次数分布のベキ指数 γ は近い値（1.14 から 1.16）をとっている。Yamamoto, and Yamazaki (2009) にこれらの詳細が述べられている。

次節以降では、なぜ二字熟語ネットワークがスモールワールドかつスケールフリーにな

表 1: 3 冊の国語辞典から得た二字熟語ネットワークの特性値。平均経路長 l が小さく、クラスター係数 C は同じ規模のランダムネットワークでの値 C_{random} よりもはるかに大きい。次数分布 $P(k)$ のベキ指数 γ の値も記す。

辞書	頂点数	辺数	平均次数	経路長	クラスター係数		ベキ指数
			$\langle k \rangle$	l	C	C_{random}	γ
広辞苑	5458	74617	27.3	3.14	0.138	0.00501	1.14
岩国	3904	32150	16.5	3.31	0.085	0.00424	1.16
三国	3444	28358	16.5	3.32	0.086	0.00483	1.14

[2]自己ループおよび多重辺の分だけ、辺の数は熟語の数より少ない

るのかを考察する。二字熟語の形成を単純化して数理モデルを立てて、モデルが生成するネットワークと実際のネットワークを定量的に比較する。

3　二字熟語ネットワークの特徴を再現するシンプルなモデル

『広辞苑』の二字熟語ネットワークにおいて次数が大きい漢字を順に挙げると大、人、水、山、子、… となっている。これら上位の漢字は、日常で頻繁に使われる重要度の高い漢字である。一方で、ごく限られた特殊な場面でしか使われない漢字も数多くある（ただし、日常生活でそういった漢字に出会うことはほとんどない）。このように、漢字のもつ重要度には大きな格差がある[3]。漢字の重要度は熟語の「つくりやすさ」に直結すると考えるのは自然な仮定である。

そこで、以下では**重要度の高い字ほど熟語をつくりやすい**という特徴を反映した数理モデルを構築する。まずは「重要度」を数値化する必要がある。本研究では漢字の「重要度」のデータとして、文化庁がおこなった『漢字出現頻度数調査 (3)』(2007) を用いた。この調査は出版物の組版データから漢字ごとの使用頻度数を算出したもので、のべ 49072315 の漢字、8576 種類の異なる漢字からなる。「人」が最も大きな出現頻度数 610660 で、一、日、大、年、… がそれに続く。モデルの説明のため、それぞれの漢字に $i = 1, 2, \ldots, 8576$ の番号をつけ、i 番目の漢字の出現頻度数を x_i とする。「漢字 i の重要度が x_i である」と解釈する[4]。

漢字 i と j が辺で結ばれるかどうか（つまり、熟語をつくるかどうか）を確率的に決める。$p_{ij} = cx_ix_j$ という確率で i と j を辺で結ぶことにする（c は定数）。例えば、$c = 10^{-6}$ の場合、$x_i = 1000, x_j = 100$ だとすると、$p_{ij} = 0.1$ であり、頂点 i と j は確率 10% で辺を結び、90% で辺を結ばないことになる。すべての頂点の組に対して辺を結ぶかどうかを判定し、ネットワークをつくっていく。出現頻度数 x_i が大きいほど、他の頂点と辺でつながる確率が高くなっている。

定数 c の値は、モデルが実際の二字熟語ネットワークに近くなるように定める。$c = 0$ のときは辺が全く張られず、すべての頂点がばらばらのネットワークができる。c を充分に大きくすると、どの 2 頂点も辺で結ばれたネットワーク（完全グラフ）ができる（図 2）。つまり、c が大きくなると「辺の密度」も大きくなることが分かる。ネットワークの辺の密度を表すのが**平均次数**$\langle k \rangle$ である。そこで、c を変化させていき、モデルの平均次数が実際の二字熟語ネットワークと一致するような c の値を探す。ここではモデルを広辞苑のネッ

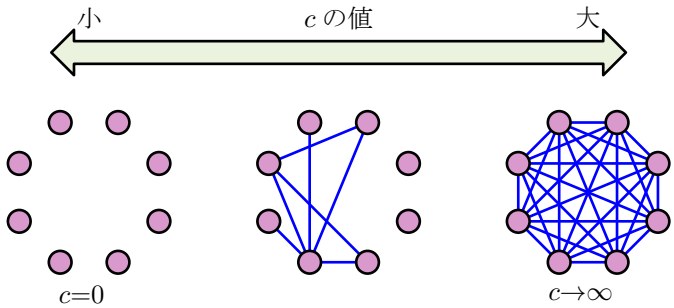

図 2: c の変化によるネットワークの構造変化のイメージ。$c = 0$ のときは辺が張られない（左）。c が充分大きいときは完全グラフとなる（右）。

[3]教育漢字、常用漢字、人名用漢字などの分類は重要度の階層を表しているといえる
[4]抽象的な「重要度の高さ」を「出現頻度数」という数値で定量化している

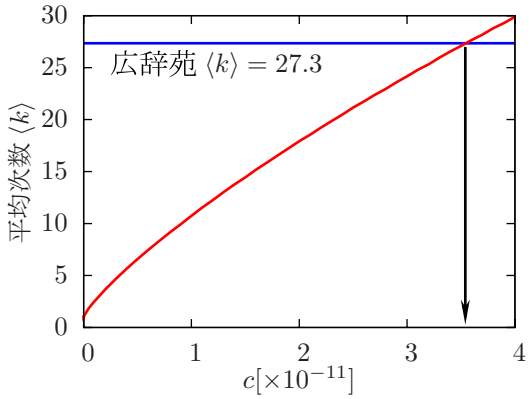

図 3: 確率 p_{ij} の比例定数 c の決定。実線はモデルの数値計算結果、破線は広辞苑のネットワークの値 $\langle k \rangle = 27.3$ である。2 つのグラフの交点から $c = 3.54 \times 10^{-11}$ が得られる。

トワークと比較することにする。c の値に対する平均次数の変化を図 3 に示す。広辞苑のネットワークの平均次数 $\langle k \rangle = 27.3$ と一致するのは $c = 3.54 \times 10^{-11}$ のときであることが分かった。

4　モデルの計算結果

前節で提案したモデルによるネットワークと実際の二字熟語ネットワーク（広辞苑）を比較したのが表 2 である。モデルには確率的な要素が入っているので、試行のたびに異なるネットワークが生成される。表に示した計算結果は 1000 個のネットワークについての平均である。『漢字出現頻度数調査 (3)』では 8576 種類の漢字が挙げられているのに対して、モデルの計算結果では頂点数が半分以上減少している。この原因は、出現頻度が小さな漢字は 1 本でも辺を獲得することが難しく、ネットワークから孤立してしまうためである（孤立した頂点は除外している）。モデルの平均次数 $\langle k \rangle$ が広辞苑と一致しているのは c の決め

表 2: モデルによるネットワークと広辞苑の二字熟語ネットワークの比較。両者の平均経路長 l, クラスター係数 C, およびベキ指数 γ はきわめて近い値である。

	頂点数	辺数	平均次数 $\langle k \rangle$	経路長 l	クラスター係数 C	ベキ指数 γ
モデル	3012	41130	27.3	2.85	0.191	1.12
広辞苑	5458	74617	27.3	3.14	0.138	1.14

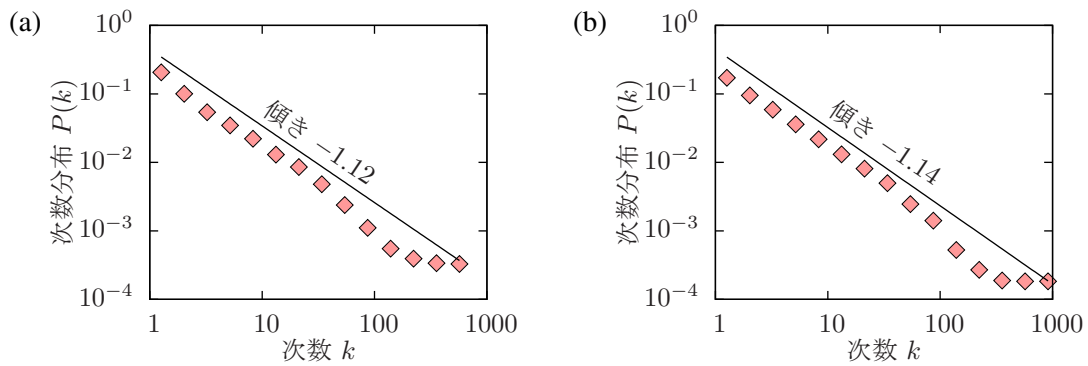

図 4: 次数分布の比較。(a) モデルが生成するネットワークの次数分布。(b) 実際の二字熟語ネットワーク（広辞苑）の次数分布。

方から当然であるが、平均経路長 l, クラスター係数 C, および次数分布のベキ指数 γ もかなり近い値となっていることは注目すべき結果である。他方、頂点数および辺数はともに“モデル”の方が“広辞苑”に比べて小さい。モデルが生成するネットワークは実際のネットワークのミニチュアであると解釈するのがよいだろう。モデルと広辞苑のネットワークの次数分布を図 4 に示す。両者はきわめて似かよっている。

5 検討

Peng, Minett, and Wang (2008) は、中国語の熟語ネットワークの構造を解析した。中国語のネットワークの経路長およびクラスター係数は日本語と近い値であるが、次数分布の指数の値は普通話および広東語でそれぞれ $\gamma = 1.40, 1.49$ となっていて日本語よりも大きい値をとる。γ が小さいほど、少数の頂点に大量の辺が集まる（次数の大きな頂点が出現しやすい）。つまり、日本語の熟語は中国語の熟語よりも重要な漢字に集中する傾向にあり、重要な漢字とそうでない漢字の格差が大きいといえる。

今回のモデルは非常に大雑把なものである。例えば、モデルでは熟語における漢字の順序を定めることができない。また、実際に二字熟語が形成する際には 2 つの漢字の意味や機能など様々な条件が考慮されるはずであるが、モデルでは熟語の形成において漢字の重要度しか考慮していない。本稿のモデルに付加的に条件を課すことで、より詳細な分析に堪えうるネットワークをつくることができるだろう。ただし、スモールワールド性およびスケールフリー性という特徴を再現するためには、本稿のような単純な設定だけで充分であることを強調する。

熟語ネットワークのスケールフリー性の起源として、Peng, Minett, and Wang (2008) は優先的接続の効果を推測している。優先的接続のモデルは、ネットワークのスケールフリー性について最初に論じた Barabási, and Albert (1999) によって提案された。成長するネットワーク（新たな頂点が次々に加わる）において、次数が大きな頂点ほど新たな頂点から辺を受け取りやすいというモデルである[5]。熟語ネットワークの形成過程で優先的接続の効果が作用したのか検証するためには、各々の熟語が「いつ形成されたのか」（各々の漢字が熟語ネットワークに「いつ加わったのか」）を知らねばならない。しかし、熟語が形成された時期の特定は困難であると考えられるので、この仮説は定性的な検証でさえ容易ではないであろう[6]。本研究のモデルではネットワークの成長を取り上げるのは避け、かわりに各漢字に内在する「重要度」に基づいて熟語のつくられ方を規定した[7]。漢字の重要度を表す数値として出版物における出現頻度の調査結果を利用し、モデルと実際のネットワークの定量的な比較をおこなうことができた。

さらに、本稿のモデルに立脚すると、日本語と中国語で γ の値が異なっているのは漢字の「重要度」の分布の仕方が両言語で異なっているためであることが導かれる。中国語における漢字の使用状況の統計を利用してこのことを確かめれば、本研究のモデルの信頼性を評価することも可能であるだろう。

---

[5]「金持ちがさらに金持ちになる (The rich gets richer.)」とたとえられる

[6]Barabási らのモデルから導かれるのは $\gamma = 3$, これを拡張した Dorogovtsev, Mendes, and Samukhin (2000) のモデルにより得られるのは $2 < \gamma < \infty$ である。中国語のネットワークのスケールフリー指数 $\gamma = 1.40$ および 1.49 がこの範囲に入らないことから、熟語ネットワークのスケールフリー性を優先的接続で説明しようとするのはそもそも的外れであると思われる

[7]本研究のモデルはフィットネスモデル (Caldarelli, Capocci, De Los Rios, and Muñoz (2002)) とよばれるモデルの一種といえる

6 まとめ

本研究では、複雑ネットワーク解析の手法を用いて二字熟語の統計的性質を調べた。漢字を頂点、二字熟語を辺とみなすことで、二字熟語は自然にネットワークの形で表現される。3 冊の国語辞典から構成したネットワークはいずれもスモールワールドかつスケールフリーであることが分かった。ネットワークの特徴を定量的に説明するために、重要度が高い漢字ほど二字熟語をつくりやすいという特徴を反映したモデルを構築した。その際、漢字の重要度のデータとして、出版物における漢字の出現頻度数の調査結果を利用した。2 つの漢字の重要度の積に比例する確率で辺を張るというシンプルな設定のモデルであるが、実際のネットワークと近い構造を再現することができた。

謝辞

漢字出現頻度数調査の冊子を提供してくださった武田康宏氏（文化庁）に感謝いたします。

文献

Duncan J. Watts, and Steven H. Strogatz (1998) “Collective dynamics of ‘small-world’ networks”, *Nature* 393, pp. 440–442.

Albert-László Barabási, and Réka Albert (1999) “Emergence of scaling in random networks”, *Science* 286, pp. 509–512.

増田直紀、今野紀雄 (2010)『複雑ネットワーク—基礎から応用まで』近代科学社.

矢久保孝介 (2013)『複雑ネットワークとその構造（連携する数学 4)』共立.

Mark E. J. Newman (2010) *Networks: An Introduction*, Oxford University Press.

Ken Yamamoto, and Yoshihiro Yamazaki (2009) “A network of two-Chinese-character compound words in the Japanese language”, *Physica A* 388:12, pp. 2555–2560.

新村出 編 (1991)『広辞苑　第四版』岩波.

西尾実、岩淵悦太郎、水谷静夫 編 (1992)『岩波国語辞典　第五版』岩波.

見坊豪紀、金田一京助、金田一晴彦、他 編 (1992)『三省堂国語辞典　第 4 版』三省堂.

文化庁 (2007)『漢字出現頻度数調査 (3)』.

Gang Peng, James W. Minett, and William S.-Y. Wang (2008) “The networks of syllables and characters in Chinese”, *J. Quant. Ling.* 15:3, pp. 243–355.

Sergey N. Dorogovtsev, José F. F. Mendes, and Alexander N. Samukhin (2000) “Structure of growing networks with preferential linking”, *Phys. Rev. Lett.* 85:21, pp. 4633–4636.

Guido Caldarelli, Andrea Capocci, Paolo De Los Rios, and Miguel A. Muñoz (2002) “Scale-free networks from varying vertex intrinsic fitness”, *Phys. Rev. Lett.* 89:25, p. 258702 [4 pages].