

星新一作品のネットワークによる分析の試み

高橋寛恒* 松原仁

(公立はこだて未来大学)†

1 はじめに

1.1 背景

本研究は「気まぐれ人工知能プロジェクト『作家ですよ』」[1]の試みの一つとして行われているものである。このプロジェクトは、コンピュータにショートショート(厳密な定義は無いが、本プロジェクトでは8000字以内の小説のことを指す)を創作させることを目的としている。また、参考にする作家として星新一氏¹を選び、彼のような作品を作ることを目標の一つとする。

この試みは、人工知能に創作という人間の感性に関する領域を扱わせようとするものである。このような、コンピュータに感性を扱わせようとする試みは、絵画[2][3]や音楽[4]、和歌[5]などの分野で行われているが、特に日本語において、散文という形態でコンピュータに感性を扱わせる試みはほとんど存在していない。そのため、この試みを成功させるためには、日本語における自然な文書生成の方法、物語の展開・プロットの決定方法や、「星新一のような」文章とはどのようなものであるのか、などの文書生成に関わる多くの課題をクリアしなければならないと予想される。

このような課題の中でも我々は特に「星新一らしさ」とは文章のどの様な要素によって成り立っているのかという問題に着目し、星新一作品を分析することによって、星新一作品を特徴付ける要素を明らかにすることを目的として活動している。

1.2 星新一作品の分析

前述のとおり、本プロジェクトではコンピュータによる、星新一のようなショートショートの創作を目標の一つとするため、星新一作品の計量的な分析が必要不可欠である。星新一作品に対する計量的分析としては、村井ら[6]により、プロット分析、tf-idf、n-gram、 χ^2 二乗検定を用いた作品のテーマ抽出と、各テーマで現れるプロットの特徴の抽出が行われている。ここで得られた知見は、ショートショート生成時のプロット決定には効果的であると思われるが、文、あるいは文章を生成することを考えた場合には、抽象度の高い情報であると考えられる。そのため、コンピュータによる文書生成の為に、より抽象度の低い、語彙や文レベルでの星新一作品における

知見が必要であると思われる。

語彙レベルの分析としては、佐藤ら[7]によって、作品における語彙の頻度情報に基づく分析が行われているが、ここで語彙としているのは名詞のみであり、その他の要素についての検討は行われていない。一般に、文章は名詞のみで構成されるわけではないので、ここで得られた知見だけでは文書生成は難しいと予想される。また、仮に動詞など他の要素についての頻度情報が得られたとしても、それらがどのように接続されて文が構成されているかという、接続関係が明らかにならなければ、文書生成にその知見を応用する事は困難であると考えられる。

以上の点を踏まえて、我々は、コンピュータによる星新一のようなショートショート生成のための知見獲得の試みとして、星新一作品のネットワークによる分析を行った。ネットワークはノード(頂点)とエッジ(辺)によって、あるものとあるものの関係を表現することが可能な知識表現の形態であると捉える事が出来る。従って、ネットワークが表現する関係を適切に設定すれば、作品に出現する語彙の接続関係を保持したまま、星新一作品における語彙の計量的情報を表現することが可能である。

本研究では、星新一作品のネットワークによる分析の第一歩として、星新一作品を元に構成したネットワークに対して分析を行い、ネットワークによる分析が、作品の分析に対して有効であるかを検証する。

2 ネットワークの構成

本節では、分析のために構成したネットワークの構成法や分析対象の設定について述べる。本研究において構成したネットワークは2種類あり、それぞれ共起ネットワークと係り受けネットワークと名づけた。続く2.1項では共起ネットワークの構成方法について述べ、2.2項では係り受けネットワークの構成方法について述べる。ネットワークによる分析対象としては、葉がテーマであると思われる32編について分析対象とした。これは、ネットワークによる分析が、作品の分析方法として有効であるかについて不明瞭であることを考慮して、適度な作品数で分析を行うためである。

作品のテーマ抽出については前述のとおり、村井ら[6]が行っているものの、ここにおけるテーマとは、作品に登場する特定の単語群であり、テーマが作品中に出現する語に限定されている点で、特に文学作品のテーマの特定方法としては疑問が残る。しかし、その他に有力な方法が確立されているわけではないため、本研究では、人手により作品のテーマを分類した。

* g2114019@fun.ac.jp

† 北海道函館市亀田中野町 116 番地 2 公立はこだて未来大学大学院

¹ 日本の作家。ショートショートと呼ばれる短編小説を生涯で1000作品以上残した。作品数の多さと、その質の高さから「ショートショートの神様」と称される。

2.1 共起ネットワーク

本項では、共起ネットワークの構成法について述べ、実際に構成した共起ネットワークを示す。共起ネットワークは、作品に出現する単語の共起関係を表現するものである。ここでの共起関係とは、一文において定義されるものであり、その文の中で出現する名詞と動詞について共起関係にあると定義する。このネットワークにおいて、ノードは作品中に存在する全ての名詞と動詞を表し、エッジは共起関係にあるノードのペアを結ぶ様に作られる。

ネットワークを構成する際には、動詞や名詞の同定に MeCab に IPA 辞書を適用したものをを用いた。この時、特定された動詞が活用形であった場合でもそれを原形に直すようなことはしていない。これは、文章生成を考える上で、動詞が活用している、活用していないという情報も重要であると考えられることによる。Fig. 1 に実際に構成した共起ネットワークを示す。以降に示すネットワークの可視化では、次数の低いノードに接続しているエッジほど青くなり、次数が高くなるほどに、緑、赤になっている。

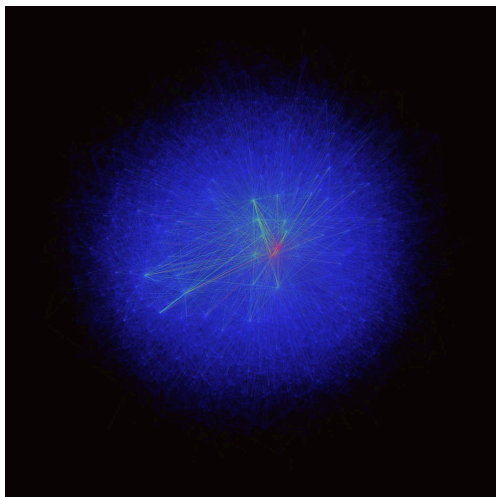


Fig. 1 共起ネットワークの可視化

2.2 係り受けネットワーク

本項では、係り受けネットワークの構成法について述べ、実際に構成した係り受けネットワークを示す。係り受けネットワークは、作品中に存在する各文の文節間の修飾関係を表すものである。文節や修飾関係の同定には CaboCha を用いた。このネットワークにおいてノードは作品の各文を構成する文節を表し、エッジは CaboCha によって同定される係り受け関係に従うようにノード間を結ぶ。この時、係る側と係られる側の関係を表すため、エッジを有向とし、係る側から係られる側に向かってエッジが張られる。Fig. 2 に実際に構成した係り受けネットワークを示す。

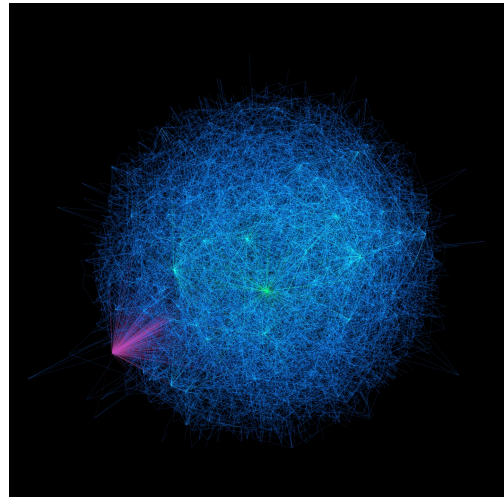


Fig. 2 係り受けネットワークの可視化

3 ネットワークによる分析

本節では、構成したネットワークによる分析結果について述べる。3.1 項では、構成したネットワークの性質を明らかにするため、2つのネットワークに対してグラフ理論に基づく分析を行った結果について述べる。次に、3.2, 3.3 項では特徴的な結果が得られた分析結果について述べる。

3.1 ネットワークの性質

共起ネットワークと係り受けネットワークの性質を調べるため、Table 1 に示すような、グラフ理論で用いられる指標を算出した。

Table 1 各ネットワークにおける指標値

指標	ネットワーク	
	共起	係り受け
ノード数	4336	11757
エッジ数	55281	1608
平均次数	25.499	1.362
平均クラスタ係数	0.748	0.003
平均パス長	2.551	6.567

また、それぞれのネットワークの次数分布を Fig. 3, 4, 5 に示す。これらのグラフにおいて、縦軸は頻度、横軸は次数を表しており、それぞれの軸は対数軸である。グラフ中では青点で累積確率分布として表した次数分布をプロットした。また、赤の実線で Aaron ら [8] の方法によるフィッティング結果を示している。係り受けネットワークについては、次数が入次数と出次数に区別できるため別々にプロットした。

これらのグラフより、ネットワークの次数分布がべき分布に従っているということがわかる。ネットワーク分析の分野では、この様な次数分布を持つネットワークをスケールフリーネットワークと呼ぶ。スケールフリーネットワークにはハブと呼ばれる非常に大きな次数を持つノード

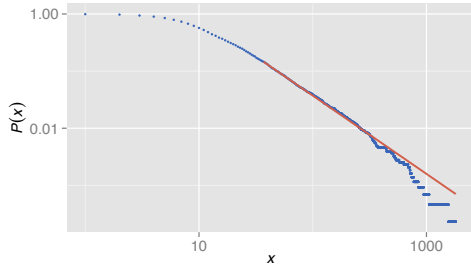


Fig. 3 共起ネットワークの次数分布

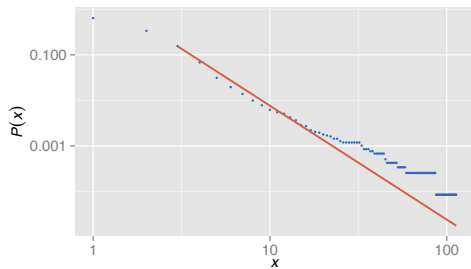


Fig. 4 係り受けネットワークの入次数分布

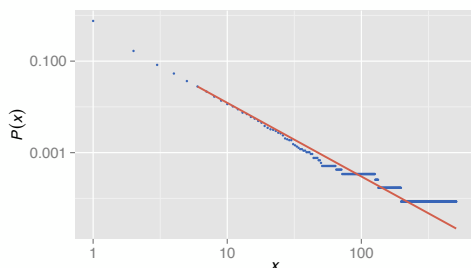


Fig. 5 係り受けネットワークの出次数分布

ドや、比較的高い次数を持つノードを持つことが知られている。共起ネットワークでは、第 3.2 項で述べるように、このようなノードに着目することで、星新一作品の特徴的語彙や、テーマに関連すると思われる語が観察できた。

3.2 共起ネットワークにおける特定ノード群への着目

本項では、共起ネットワークにおいてある特定のノード群について着目した結果について述べる。ここでの特定ノード群とは、ノードを次数順に全体の上位 5% にあたる数まで取り出したものである。これらのノードに着目した結果、星新一作品によく見られると言われる特徴的な語彙や、テーマに関連すると思われる語を観察することが出来た。観察された語について一部を抜粋したものを Table 2 に示す。

星新一は作品の中で特徴的な人物表現²を用いるということが知られているが、これらの特定ノード群の中に

² エヌ・エフ・彼女などの表現を多用し、特定の人名を用いることを避ける。

Table 2 観察された語彙 (抜粋)

観察されたノード	次数
エヌ	374
エフ	247
男	319
青年	213
彼女	184
博士	513
薬	714
作用	225
薬品	191
研究所	122
効果	143

もエヌやエフ、男や青年、彼女や博士などといった人物表現が見られた。また、本研究ではテーマが薬に関連すると思われる作品のみを集めて分析を行っているため、薬、作用、薬品、研究所、効果などテーマに関連すると思われる語が見られた。

このような語が次数に着目することで発見可能であることは、ネットワークの構成法と関連があるものと思われる。このネットワークのエッジは、前述した共起関係にあるノード間を全て結ぶので、ある 1 つの文についての部分グラフは、完全グラフとなる。これは、繰り返し出現する単語のノードほど、多くのノードと隣接しやすくなることを意味する。これにより、次数に着目することで、作品で多用される人物表現や、テーマに関連すると思われるような語彙が抽出可能になったものと考えられる。

3.3 係り受けネットワークにおける生成へのランダムウォーク適用

本項では、係り受けネットワークに対して適用した、ランダムウォークによる短文生成の試みについて述べる。ここでのランダムウォークとは、ネットワーク上のノードを移動するような仮想的なエージェントを考え、そのエージェントをある停止条件になるまでネットワーク上で動作させることを指す。本研究で行ったランダムウォークの手順は以下のようなものである。

1. エージェントの初期位置となるノードをランダムに選ぶ
2. エージェントの現在地ノードから出ているエッジの先のノードにランダムに移動する
3. 2 を次のいずれかになるまで続ける
 - (a) エージェントの移動したノード総数がある数に達する
 - (b) 現在のノードから出ているエッジが存在しないノードに到達する (出次数が 0 のノード)

以上のような手順でランダムウォークを行い、エージェントが到達した順番にノードが表す文節を繋げていった結果、Table 3 に示すような短文が得られた。

Table 3 生成された文 (抜粋)

生成された文	文節数
崇高な使命なのだという妄想が、発生したのだろう。	4
乱すのは人道上しのびないが、遂行しなければならない。	3
宝物を当てた、といった感想を抱く人も多い	7
毛髪のはうはなんとかなる。	3
お金さえあれば負けずに働けます。	4
ネコは魔物というから、やりかねない。	4

この例では、(a)の条件を、7個のノードを經由したら終了としてランダムウォークを実行した。この方法はランダム性が取り入れられているため、必ず意味の通る文が生成されるわけでは無いが、Table 3 に示した短文は、一見して意味が通ると判断され、かつ、ネットワークを構成する際に使用した作品の中には存在しない、新規の文である。

この方法の問題点としては、生成のされる文が元の作品に出現した文節の修飾関係に依存する為、新規の文であっても、元の作品と類似した文ができてしまうことが挙げられる。このことは、星新一のような作品を作る上では好ましいと考える事もできるが、同時に剽窃に近づくということでもある。従って、星新一らしさを大きく損なわない範囲で剽窃から脱却する方法を模索する必要がある。また、この方法はあくまでも短文生成のための方法であり、複数の文をまとめて一つの文章にするようなことは出来ない。そのため、文章生成に対する直接の貢献にはならないが、既存の作品から意味の通る新規の文が生成できたという点で、星新一らしい文章生成の第一歩になるものと考えている。

4 まとめ

本研究では、コンピュータによる星新一のようなショートショート生成のための知見獲得の試みとして、星新一作品のネットワークによる分析を行った。分析では、特に薬をテーマにしていると思われる、32編を対象に、共起ネットワークと係り受けネットワークの2つのネットワークを構成した。

その結果、共起ネットワークにおいては、特定ノード群に着目することで、星新一作品における特徴的な人物表現や、テーマに関連すると思われる語を観察することが出来た。係り受けネットワークでは、ランダムウォークを適用し、既存の星新一作品から短文を生成する方法について示した。以上のことから、ネットワークによる分析は、生成に向けた作品の分析方法として有効であると考えられる。

今回の試みは実験的なものであり、星新一のような文書(作品)生成という目標に対して、十分な知見を提供できたとは考えていないが、その最初の段階として、ネットワーク分析の有効性を示せたものと考えている。また、頻度分析には無いものとして、語や文節間の接続情報が

ネットワークでは表現されているため、これらの情報が文書生成に役立つことを期待している。

今後の展望として、本研究で対象としたものよりも大規模な作品群での適用や、名詞と動詞以外の要素についても考慮したネットワークの分析などを考えている。また、今回の分析で用いた手法を、他作者の作品に適用することで、星新一作品と比較して、どのようなところが異なっているかについても探っていきたい。

参考文献

- [1] 松原 仁, 佐藤理史, 赤石美奈, 角 薫, 迎山和司, 中島秀之, 瀬名秀明, 村井 源, 大塚裕子: コンピュータに星新一のようなショートショートを創作させる試み, 2013 年度人工知能学会全国大会論文集, <https://kaigi.org/jsai/webprogram/2013/paper-110.html>, 人工知能学会 (2013).
- [2] Kurzweil CyberArt Technologies Inc.: AARON: A Product of Kurzweil CyberArt Technologies, <http://www.kurzweilcyberart.com/>.
- [3] 迎山和司: 人工知能画家 静第3版 (ペン入力インタフェース (2)), 情報処理学会研究報告. HI, ヒューマンインタフェース研究会報告, Vol.2004, No.51, pp. 47-54 (オンライン), <http://ci.nii.ac.jp/naid/110002911832/> (2004).
- [4] 深山 覚, 中妻 啓, 米林裕一郎, 酒向慎司, 西本卓也, 小野順貴, 嵯峨山茂樹: Orpheus : 歌詞の韻律に基づいた自動作曲システム (自動作曲), 情報処理学会研究報告. [音楽情報科学], Vol.2008, No.78, pp. 179-184 (オンライン), <http://ci.nii.ac.jp/naid/110006885356/> (2008).
- [5] 佐々木 あらら: 星野しずるの犬猿短歌, <http://www17.atpages.jp/sasakiarara/sizzle/>.
- [6] 村井 源, 松本斉子, 佐藤知恵, 往住彰文: 物語構造の計量分析に向けて: 星新一のショートショートの物語構造の特徴, 情報知識学会誌, Vol.21, No.1, pp. 6-17 (オンライン), <http://ci.nii.ac.jp/naid/10029134488/> (2011).
- [7] 佐藤知恵, 村井 源, 往住彰文: 文学作品群の特徴的語彙と概念カテゴリーの抽出: 星新一ショートショートの計量分析 (特集 第17回 (2009年度) 年次大会 (研究報告会&総会)), 情報知識学会誌, Vol.19, No.2, pp. 132-137 (オンライン), <http://ci.nii.ac.jp/naid/110007227673/> (2009).
- [8] Aaron, C., Cosma Rohilla, S. and M. E. J. Newman.: Power-Law Distributions in Empirical Data, *SIAM Rev.*, Vol.51, No.4, pp. 661-703 (online), <http://dx.doi.org/10.1137/070710111> (2009).