Vol.2019-CH-119 No.10
2019/2/16

19世紀の推理小説：機械学習アプローチによる文体分析

黒田 絢香[†1,a)]

概要：本稿は，Arthur Conan Doyle の作品を中心として構築した同時代の推理小説コーパスを分析対象とし，機械学習手法の一つであるトピックモデル (LDA) を用いて，推理小説というジャンルにおける言語的特徴を探るものである．計量的・統計的なアプローチから作品ごと，あるいは作家ごとの特徴語を抽出し，作品群がどのようなトピックを持っているか，周辺作品との共通点や相違点，関係性がどのようなものか論じる．

キーワード：推理小説，トピックモデル，LDA，文体分析

1. はじめに

デジタル・ヒューマニティーズと呼ばれる人文学研究と情報科学の中間に位置する学際融合的な分野では，人文学における諸問題を解決するアプローチとして様々なデジタル技術や計量的・統計的手法が提案されている．特に文学研究においては，著者推定や文体論，大規模テクスト分析など研究課題に応用され，従来の人手による精読的なアプローチとは違った観点から作品を見つめ直すことができるようになった．

本稿では，この技術を 19 世紀ごろの推理小説に適用し，このジャンルの代表的作家である Arthur Conan Doyle の作品を中心として分析することで，推理小説に特徴的に多く使われている単語や作品傾向，作者ごとの言語的特徴の差異を検討することを目標とする．これまで人文学の分野で行われてきた分析とは違った観点から光を当てることにより，従来の研究結果を裏付ける客観的な証拠，あるいは通説にはなかった事実を発見し，推理小説分析をさらに深めることができると考えられる．

2. 推理小説の歴史と意義

殺人や盗難などの事件を探偵，あるいは警察が解決する過程を描く推理小説 (Detective fiction) は，古くから多くの人々に愛される小説ジャンルの一つである．推理小説の始祖は 1841 年にアメリカで発表された Edgar Allan Poe の *The Murders in the Rue Morgue*(『モルグ街の殺人』) であると言われているが [1][2]，このジャンルを広く世に知らしめたのはイギリスにて 1887 年より発表された Arthur Conan Doyle の Sherlock Holmes シリーズである．

それまでの推理小説は短編小説が中心であったが, 1913 年の長編推理小説である *Trent's Last Case* が Edmund Clerihew Bentley によって執筆されて以降，Agatha Christie や Ellery Queen など様々な作家が，魅力的な探偵の登場する本格的な長編を次々と発表する，黄金時代と呼ばれる時代に至った [2]．それ以後現在に至るまで，推理小説は国を問わず様々な地で生み出され，また小説に限らず映画や漫画，ゲームなど多様なメディアで展開される “ミステリ” という一大ジャンルになった．

推理小説の誕生の背景には，当時の社会情勢が大きく関わっている．推理小説は新聞の犯罪報道のように，民衆の犯罪に対する興味と不安をわかりやすく表現した上でその不安を解消するものであり，娯楽として消費されると同時に，近代社会の社会性を再確認する機能があった [3]．また，1830 年代にイギリスの警察制度が整備されたことも大きく関係していると考えられる．すなわち，公権力の監視機能への大きな関心とその不信感が，公人ではなく私人として事件を解決する『探偵』という存在のヒーロー視へと繋がったということである．

つまり，推理小説の分析は文学作品の一ジャンルの分析という点にとどまらず，執筆当時の社会状況，公権力と民衆の関係性，犯罪や司法のあり方，近代社会における意識の変化を捉えるという点で有意義であると言える．

[†1] 現在，大阪大学大学院言語文化研究科
Presently with Graduate School of Language and Culture, University of Osaka
[a)] kuroda22a@gmail.com

3. 文体分析とトピックモデル

3.1 テクストの『キーワード』とは何か

計量的なテクスト分析において，キーワード (key-word)，もしくは特徴語を特定することは重要なプロセスの一つである．多くの場合，キーワードはその文書の著者や読者によって主観的に決められるが，コーパス言語学の文脈では，統計学の考え方に基づいてこれを客観的な数値データから抽出する試みが数多くなされている．分析対象の言語的特徴を集約する単語を機械的に選び出すことで，大規模なデータを効果的に要約することができるためである．

Wynne[4] は key-word の定義について，“words which can be shown to occur in the text with a frequency greater than the expected frequency (using some relevant measure), to an extent which is statistically significant”，つまりある文書において期待値よりも統計的に有意に多く出現する語であると述べている．ある特定の単語に注目し，文書間，あるいは文書集合間でその頻度に統計的に有意な差があるかどうかを検証することでキーワードを特定する．

有意性を測るため，これまで χ^2 検定や対数尤度比検定，Mann-Whitney の U 検定など様々な統計的指標が提案され，テクスト分析に大きく貢献してきた．代表的な研究の一つとして挙げられるのは，アメリカ英語とイギリス英語の言語的特徴を分析するため，アメリカ英語の書き言葉を収録した Brown コーパス，及び同様の構成でイギリス英語を収録した Lancaster-Oslo/Bergen (LOB) コーパスを χ^2 検定を指標として比較した研究である [5][6].

これらには一方で，手法によって限界も存在する．例えば χ^2 検定や対数尤度比検定では，抽出したキーワードがコーパス内でどのように分布しているかを考慮できない．そのため，特に長編小説を含む文学作品コーパスでは，一作品にのみ突出して多く出現する単語をコーパス全体の特徴語とみなしてしまう問題が発生しうる．Mann-Whitney の U 検定ではこの問題をカバーし，コーパス内で分布に一貫性があるかどうかを評価指標に組み込んでいるが，この場合は作品ごとのローカルな特徴を取りこぼしてしまう．また，χ^2 検定では高頻度語，U 検定では低頻度語にそれぞれ重点を置く傾向があるため，抽出したい語の傾向に従って慎重に指標を選択する必要がある．そして，これらの手法から得られるキーワードは，時に何百語，何千語の羅列となってしまい，特徴語リストの背後にある意味的な構造を解釈するために多大な労力がかかってしまうことがある．

3.2 トピックモデルとテクスト分析

そこで本研究は，機械学習のアルゴリズムであるトピックモデルに注目する．大規模なテクストの自動分類を目的として開発されたこのモデルは，「各単語は潜在的にトピック (話題) を持ち，同じトピックを持つ単語は同一の文書に出現しやすい」という想定を前提とし，一つの文書に共起しやすい単語の集合を「トピック」として抽出する．例えば，「野球」という単語が出現する文書には，「選手」や「チーム」，「決勝」，「ホームラン」などの語が共起する確率が高く，一方で「選挙」が含まれる文書においては前述の単語の出現確率は下がり，「投票」，「有権者」，「議員」，「国政」などが出現しやすくなる．このように同じ文書に出現しやすい語のグループをトピックと呼び，前者のグループが多く含まれる文書はスポーツをテーマとして，後者が多く含まれる文書は政治をテーマとして書かれたものであると判断することができる．

このモデルの特徴の一つに，教師なし学習であることが挙げられる．つまり，各単語の意味をあらかじめラベル付けしたり，膨大な辞書データを参照したりする必要がなく，ただ単語の頻度分布だけで文書のテーマを予測することができる．

あるテクストについて，どのようなトピックが多く出現しているか，そのトピックはどのような語で形成されているかを観察することで，分析対象のテクスト，そしてコーパス全体の意味的な特徴を俯瞰できる．これは，前述した統計手法の限界の一つである，意味的構造の解釈に大きな手間がかかってしまうという点を解決することができると考えられている．

3.3 潜在的ディリクレ配分モデル (LDA)

本研究では，トピックモデルの一つである潜在的ディリクレ配分モデル (Latent Dirichlet Allocation) を用いる．Blei ら [7] らによって提案されたこのモデルの特徴は，一つの文書に複数のトピックが潜在していることを前提としていることである．

以下の図 1 は単語 w，トピック z，トピック分布 θ，単語分布 ϕ，ハイパーパラメータ α，β の関係性を表すグラフィカルモデルである．外側の四角は各文書に N 個の単語，文書集合に D 個の文書が含まれていることを示しており，K はトピック数である．つまり，トピックごとに異なる単語分布 ϕ が，文書ごとに異なるトピック分布 θ が存在している．

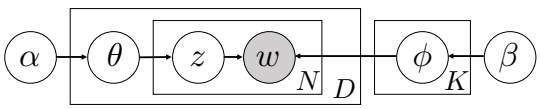

図 1 LDA のグラフィカルモデル

LDA において算出されるのは，「文書ごとのトピック出現確率の分布」と「トピックごとの単語出現確率の分布」である．この二つの確率分布を読み解くことで，文書ごとに特徴的に出現する単語を発見できるのみならず，同じト

ピックに出現する意味的に近い語，似たトピック構造を持つ文書など，単語間，文書間の複雑な関係性を観察することができる．

LDA をテクスト分析に用いた例として，Jockers ら [8] による 19 世紀のフィクション作品を集めた大規模コーパスから男性作家，あるいは女性作家に特徴的に多く使用される 25 のトピックを抽出した研究が挙げられる．また，田畑 [9] は前述の二つの確率分布をより分かりやすく捉えるため，ネットワークグラフやヒートマップなどの視覚化手法を提案しながら，FLOB コーパスの背後にある意味的構造を明らかにした．

4. 手法

4.1 データと前処理

本研究の主な分析対象は前述の通り，推理小説の代表的な作家の一人である Arthur Conan Doyle である．そして比較対象として，Doyle よりも以前に作品を発表していた Edgar Allan Poe，後に発表した Agatha Christie を選出した．彼らは作品を発表した時期こそずれているものの同時代を生きた作家であり，現在でも推理小説の巨匠として広く名を知られている作家ある．

Doyle と Christie は短編，長編ともに数多くの作品を発表しているが，一方で Poe の作品群には推理小説と呼ばれるものが少ない．一般的に推理小説として受容されているのは *The Murders in the Rue Morgue* を始めとする 3 つの短編のみで，場合によってはそれらに *The Gold-Bug* と *Thou Art the Man* も加えられる [2].

計量的な分析を行う際比較対象のデータ量に大きな差があると，得られた特徴がテクストに内在するものなのか，文章量の差によるものなのかが判別できない．本研究のように作家間の差や作品間の差に注目したい場合，それ以外の形式はなるべく揃えた方が好ましい．そのため，Poe の作品数とその単語数に合わせ，Doyle と Christie の作品の中でも推理小説の短編を分析対象とした．Doyle からは Sherlock Holmes シリーズでは一作目の短編集である *The Adventures of Sherlock Holmes* に収録された 12 作品，Christie からは Poirot シリーズの一作目 *Poirot Investigates* に収録された 11 作品を選出し，Poe の 5 作品と合わせた 28 の短編を分析する．以下の表 1 は作品のリストである．

表 1 分析対象テクスト

題名	発表年	ラベル
The Murders in the Rue Morgue	1841	Poe_1
The Mystery of Marie Roget	1842	Poe_2
The Gold-Bug	1843	Poe_3
Thou Art the Man	1844	Poe_4
The Purloined Letter	1845	Poe_5
A Scandal in Bohemia	1891	Doyle_1
The Red-Headed League	1891	Doyle_2
A Case Of Identity	1891	Doyle_3
The Boscombe Valley Mystery	1891	Doyle_4
The Five Orange Pips	1891	Doyle_5
The Man with the Twisted Lip	1891	Doyle_6
The Adventure of the Blue Carbuncle	1892	Doyle_7
The Adventure of the Speckled Band	1892	Doyle_8
The Adventure of the Engineer's Thumb	1892	Doyle_9
The Adventure of the Noble Bachelor	1892	Doyle_10
The Adventure of the Beryl Coronet	1892	Doyle_11
The Adventure of the Copper Beeches	1892	Doyle_12
The Adventure of the Western Star	1924	Christie_1
The Tragedy at Marsdon Manor	1924	Christie_2
The Adventure of the Cheap Flat	1924	Christie_3
The Mystery of Hunter's Lodge	1924	Christie_4
The Million Dollar Bond Robbery	1924	Christie_5
The Adventure of the Egyptian Tomb	1924	Christie_6
The Jewel Robbery at the Grand Metropolitan	1924	Christie_7
The Kidnapped Prime Minister	1924	Christie_8
The Disappearance of Mr Davenheim	1924	Christie_9
The Adventure of the Italian Nobleman	1924	Christie_10
The Case of the Missing Will	1924	Christie_11

収集した作品データに対して，分析の前処理として

- 目次，挿絵，注，冒頭の引用など，本文以外のものを削除
- 人名，地名などの固有名詞を削除
- 各文書を 1,000 語ごとに分割

の 3 つを行った．

固有名詞の特定には品詞タグ付けソフトウェアである Tree Tagger[*1] を用い，NP または NPS とタグ付けられたものを全て取り除いた．この処理は，特定の人名や地名が特徴語として強く重み付けされてしまい，言語的特徴の僅かな差異を見づらくしてしまうことを防ぐために行う．特に小説の場合，特定のキャラクタ名 (holmes や watson など) の頻度が極端に高いことから，この処理は非常に重要である．

テクスト分割も LDA の実装には欠かせない前処理である．前述の通り，トピックとは一つの文書に共起しやすい単語の集合であるが，分析対象の文書が長すぎる場合，一文書中で言及される話題の数が膨大になり，結果として雑多な単語の多く含まれる漠然としたトピックばかりが出力されてしまう [8]．また，今回用いた方式では頻度の相対化を行わないため，文書が長ければ長いほど相対的に各単語の出現頻度が上がり，トピックに強く重み付けされる傾向

[*1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

になる．そのため，本実験では先行研究に倣い [8]，各作品を 1,000 語ごとに分割した 202 のサンプルを分析対象とした (分割の結果 1,000 語に満たなかったファイルは切り捨てとしている)．

4.2 ソフトウェアと設定

LDA の実装には java ベースのツールキットである MALLET[*2] を用いた．このツールでは抽出するトピックの数を自由に設定できるため，これを 20 から 200 まで変化させながら実験を行った．この数値はトピックの粒度に関係し，小さすぎれば様々な単語が一挙に含まれる茫洋なトピックになってしまい，大きすぎればローカルな特徴を抽出できる一方でトピック間，単語間の複雑な関係を観察し解釈することが難しくなってしまう．分析目的やコーパスの規模によって最適解が異なるため，細かく数値を変えて実験を重ねることが重要である．本稿では，トピック数を 50 に設定した場合を中心に取り上げた．分析対象が全て推理小説であるため各文書に共通する話題が多く，トピック数を多く設定してしまうと関連する話題が過剰に分割されてしまう結果になったためである．

また，分析結果の可視化には統計解析ソフトウェアの R[*3]，及びネットワークグラフの可視化・解析ソフトウェアの Gephi[*4] を用いた．

5. 結果と考察

図 2 は，文書ごとに出現するトピックの確率分布データを元に描かれたネットワークグラフである．数字のノードはトピックを表しており，そのトピックが含まれるテクストはエッジによって繋がっている．また，エッジの太さは重み付けに対応している．

中心部分に位置し様々なテクストノードと繋がっているトピックノード，すなわち 35 や 37，25 などは，様々な作者，作品に共通して多く出現するトピックを表している．一方で，外縁部に位置し少数のノードとしか繋がっていないトピックは，そのテクストサンプルにおけるローカルな特徴であると言える．

例として，比較的中心部に位置するトピック 37 を挙げる．図 3 はトピックに含まれる単語をワードクラウドの形で表現したもので，各文字の大きさがそのトピックにおける重みに対応している．この図から，トピック 37 は room, house, door などを中心とした家・家具に関するトピックであると言える．このトピックの分布を表した図 4 を参照すると，どの作家の作品にも出現しているものの，比較的 Doyle が多く使用している一方で，Poe の作品にはあまり多く出現しないことがわかる．これは，事件の発生する舞台や調査の中心地が室内であることが多いという Doyle 作品の特徴を示唆していると考えられる．

図 3 トピック 37 に含まれる語

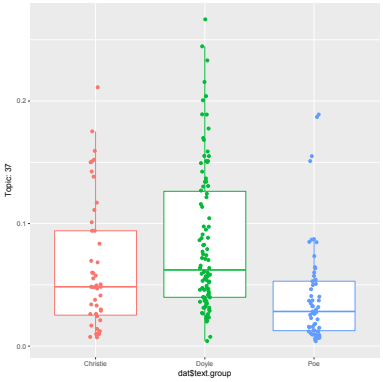

図 4 トピック 37 の作家間分布

一方で，特定の作家や作品に特徴的に出現するトピックについて着目することも有意義である．例えば，以下の図 5，6 に表されるトピック 48 は，Poe の作品に比較的多く出現する一方，他の二者の作品にはほとんど出現しない．このトピックは，body, corpse, murderers など死者や死体に関わる単語が数多く含まれており，暗号解読がメインの謎である Poe_3，“The Gold-Bug” を除くすべての Poe 作品において出現していた．Poe の作品における死体描写の特徴が反映されたトピックであると言える．

図 5 トピック 48 に含まれる語

[*2] http://mallet.cs.umass.edu/
[*3] https://www.r-project.org/
[*4] http://oss.infoscience.co.jp/gephi/gephi.org/index.html

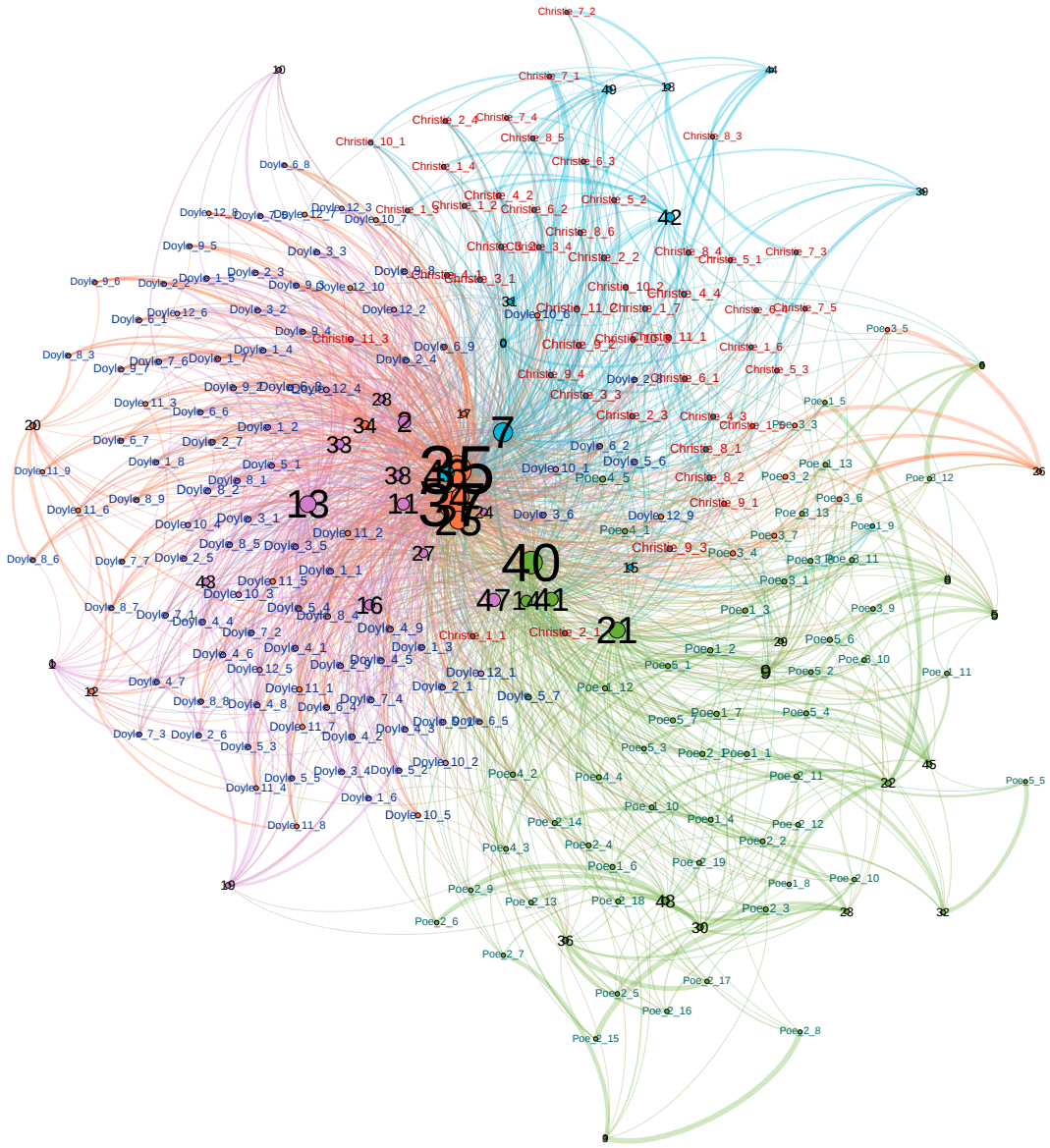

図 2　テクストとトピックの関係性を図示するネットワークグラフ

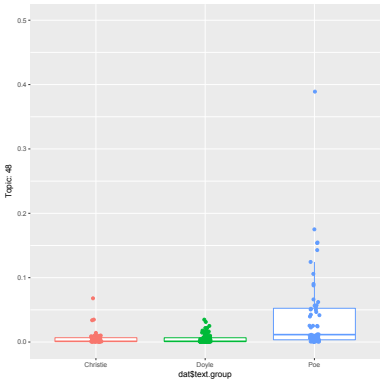

図 6　トピック 48 の作家間分布

本研究で比較した 3 作家の中で唯一の女性である Christie だが，文体に女性的な要素は反映されているのだろうか．50 のトピックのうち，Cristie 作品にのみ特徴的に出現しているトピックを幾つか発見したが，そのいずれも各作品に描かれているテーマや主人公に関係したものであった．例えば，図 7 はフランス語が多く含まれるトピックで，これはこの短編集の主人公である Hercule Poirot がフランス語混じりで話すことが関係している．

図 7　トピック 42 に含まれる語

一方で，woman や lady など女性に関係するトピックが多く出現していたトピック 2(図 8, 9) は，Christie のみならず Doyle の作品にも多く出ており，作家の性別というよりも事件の登場人物の傾向を反映している．Doyle と Christie の作品は，前述のトピック 37 にも表れているように，どちらも家庭内で起こる事件を多く取り扱っている．

作家の性別による傾向については，男女ともに作家のサンプル数を増やし再考する必要があると考えられる．

図 8　トピック 2 に含まれる語

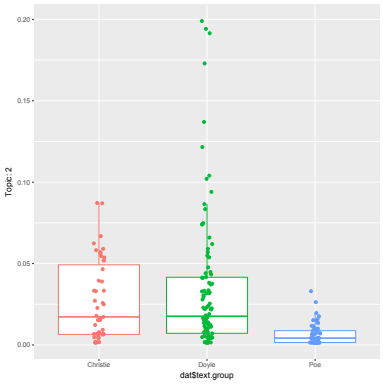

図 9　トピック 2 の作家間分布

6. まとめと今後の課題

参考文献

[1] 森秀俊：世界ミステリ作家事典 [本格派篇], 国書刊行会 (1998).
[2] 長谷川秀記：世界の推理小説・総解説，自由国民社 (1991).
[3] 内田隆三：探偵小説の社会学，岩波書店 (2011).
[4] Wynne, M.: *Searching and Concordancing*. In Kytö, M., Lüdeling, A., and Gruyter, M. (eds.) *Corpus Linguistics*.1: 706–737(2008).
[5] Hofland, K. and Johansson, S.: *Word Frequencies in British and American English*. Norwegian Computing Centre for the Humanities (1982).
[6] Leech, G. and R. Fallon.: Computer corpora—what do they tell us about culture? *ICAME Journal*, 16: 29–50 (1992).
[7] Blei, M., Ng, A. and Jordan, M.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022 (2003).
[8] Jockers, M. and Mimno, D.: Significant themes in 19th-century literature. *Poetics* 41: 750–769 (2013).
[9] 田畑智司：FLOB コーパスの意味構造: 確率論的トピックモデルによる言語使用域の特徴付け]『統計数理研究所共同研究リポート』386: 1–17 (2017).
[10] Schöch, C: Topic Modeling with MALLET: Hyperparameter Optimization(online),
入手先 ⟨http://dragonfly.hypotheses.org/1051⟩
(2019.01.24).