近現代の日本語文学作品における発表年次の予測

小川稜真　久野雅樹
電気通信大学大学院情報理工学研究科
o2430022@edu.cc.uec.ac.jp　hisano@uec.ac.jp

概要

本研究では,文学作品のテキストデータを用いた年次予測の可能性を検討した．まず，芥川龍之介の作品を対象に，TF-IDF を特徴量として複数の回帰モデルを用い，初出年の予測を行った．その後，青空文庫に収録された大量の作品に対象を拡大し，同様に年次予測を試みた．いずれの場合も，予測は一定の精度で可能であることが示され，使用される語彙のパタンが年次の推定に有効であることがわかった．特に，ランダムフォレスト回帰が他のモデルと比較して優れた予測性能を示した．

1 はじめに

文学作品の執筆時期を推定する研究は，作家の文体変化や歴史的背景を分析するうえで重要な役割を果たしている．特に，執筆年次が不明な作品や，作家の創作活動における時期ごとの作風の変化を探る研究では，文章の特徴量を定量的に分析する手法が有効とされている．こうした研究は，文学作品だけでなく，歴史的文章や手紙の分類，文章鑑定など幅広い分野で応用可能である．

金(2009)は，芥川龍之介の作品を対象として執筆時期推定を試みた[1]．形態素解析を用いて文章中の助詞や句読点の使用頻度を特徴量とし，線形回帰(LR) ランダムフォレスト回帰(RFR)やサポートベクター回帰(SVR)といった手法で分析を行った．その結果，特定の助詞の使用頻度が執筆年次に応じて変化していることが明らかとなり，特に RFR や SVR では高い予測精度が得られた．これにより，形態素レベルの文法要素が執筆時期の推定において有効な手掛かりとなり得ることが示されている．しかし，他の類似の研究を含めても，対象とする作家のバラエティや作品数，また分析に用いる指標が少ないことが課題として残されている．

本研究では，こうした課題に対応するため，芥川龍之介の作品に加え，青空文庫に収録された複数の作家の作品を対象に，執筆時期推定の可能性を検証した．分析には語彙情報を重視する TF-IDF を特徴量として用い，回帰モデルを構築した．まず，芥川龍之介の作品に実験を行い，その後，多くの作家の多くの作品を対象に同様の手法を適用した．

以上のことを通して本研究は，語彙情報を活用した執筆時期推定の有効性を示し，文体変化分析の拡張に寄与するものである．

2 関連研究

歴史的な文書や文学作品の年次推定は，文献学や自然言語処理の分野で重要な研究課題である．年次推定を行うことで，文章の歴史的背景の理解やテキスト変遷の分析が進む．

Boldsen ら(2021)は，中世スウェーデンやデンマークの公文書，英語ニュースコーパスを対象に，SVM やナイーブベイズなどの分類手法とガウス過程回帰を比較し，特に文字 n-gram を特徴量とする手法が年次推定に効果的であることを示した[2]．文字 n-gram はスペリングや形態変化を捉えやすく，歴史的文書の年代特定に有効であるとした．また，分類手法は明確な時系列データで高い精度を示し，一方で回帰手法は年代が不明確な文書において安定した結果を示した．これにより，文章の特性に応じたモデル選択の重要性が示された．

川崎ら(2022)の研究では，スペイン語古文書を対象に複数の回帰モデルを比較し，年次と地点の同時推定を行っている[3]．この研究では，加重平均 k-NN が最も高い推定精度を示し，次いでガウス過程回帰の有効性が確認した．また，文字 n-gram を特徴量としたモデルが，分散表現ベースモデル(Doc2Vec，BERT)を上回る結果を示し，文章内の特定の語や表現が年次推定において重要であることを示唆した．

小林ら(2015)らの研究では，日本の流行歌の歌詞を対象に計量文体論の手法を用いて時系列変化を分析している[4]．1977 年から 2012 年までの歌詞を対象に，品詞，語種，文字種，語彙レベルなど 26 種類の語彙指標を説明変数として重回帰分析を試みた．その結果，助動詞や連体詞，漢語，外来語の使用率が楽曲の流行年と強い相関を持つことが明らかにな

り，回帰モデルによる年代推定が有効であることが示された．

李(2022)の研究では，近現代日本語小説を対象に，助詞，文末表現，接続表現といった言語項目に着目し，経時的変化を分析している[5]．この研究では，RFR や Elastic Net 回帰などの回帰手法を用いて，特定の言語特徴が時代ごとにどのように変化しているのかを定量的に明らかにした．特に，RFR を用いた分析では，助詞の使用頻度が年代推定に寄与する重要な特徴であることを示した．これにより，テキストの特徴量を用いた年次推定モデルの構築において，統計的手法と機械学習の融合が有効であることを示した．

3 青空文庫

青空文庫は,日本の著作権保護期間が満了した作品や承諾を得た作品を無料で公開する電子図書館であり, 1997 年の設立以来，非営利で運営されている．明治，大正，昭和期の近現代の日本語の文学作品が中心に収録され，芥川龍之介や森鷗外，太宰治といった文豪の作品がテキスト形式で公開されている．近年では，自然言語処理のデータセットとしても活用されており，年代推定や著者識別の研究における重要な言語資源となっている．また，教育や研究の場においても活用されており，日本文学の分析や機械学習を活用した文学解析など幅広い分野で貢献している．

4 分析 1:芥川作品の初出年次推定

4.1 目的

先行研究(金[1])で用いられていた助詞や句読点の使用頻度を特徴量とする手法と比較し, TF-IDF を用いた語彙情報を特徴量とする手法の有効性を検証することを目的とする．具体的には，芥川龍之介の作品を対象に TF-IDF を用いた特徴量が執筆時期の推定に有効であるかを明らかにすることを目指す．

4.2 使用したデータセット

分析対象として Hugging Face[6]から取得した青空文庫収録作品のデータセットを使用した．このデータセットには，作品 ID，タイトル，作者名，テキスト，初出年などの情報が含まれ，分析対象として芥川龍之介の作品を抽出した．前処理として，テキストの分かち書きには MeCab を用いて形態素解析を行い，旧字旧仮名で記述された作品，初出年の情報がない作品，テキストが欠落している作品，重複している作品も除去した．また，初出年の年は西暦に，月は10進数に変換した．以上のような手順でデータセットを構築し，分析対象とする作品数は 192 となった．

4.3 実験手順

分析では，TF-IDF ベクトル化を用いてテキストをベクトル表現に変換した．この際，最大次元数の選択は TF-IDF スコアの高い単語を優先して選択する方法を採用し, 768 次元数に設定した．この次元数は，BERT の入力次元に合わせたものである．他の次元数による分析も行っているが，本稿では割愛する．これは分析 2 でも同様である．次に，線形回帰モデル(LR)，サポートベクター回帰(SVR)，ランダムフォレスト回帰(RFR)の 3 つのモデルを使用し，データを 5 分割交差検証法に基づき，訓練用 80%とテスト用 20%に分割して学習と評価を行った.さらにモデルの基本性能やデータ分割の影響を確認するために交差検証なしの方法でも評価を行った．各モデルにおいて，平均絶対誤差(MAE)，中央値絶対誤差(MedAE)，平方根平均二乗誤差(RMSE)，決定係数(R^2)の 4 つの指標を用いて性能を評価した．

4.4 結果と考察

表 1 の結果より, LR では MAE が 1.502 年, MedAE が 1.155 年, RMSE が 2.061 年と安定した性能を示した. SVR は全体的に誤差が大きく，特に MAE が 2.282 年, RMSE が 2.849 年, R^2 では 0.448 と他のモデルよりも誤差が大きい結果となった．一方で RFR は, MAE が 1.448 年, MedAE が 0.954 年, RMSE が 2.146 年と全体的に良好な性能を示し，特に MedAE において他のモデルより優れた結果となった．

表 2 の交差検証なしの結果では，全般的に交差検証時よりも良好な結果を示した. 3 つのモデルの優劣についても交差検証の場合とほぼ同様である．

以上の結果より，一人の作家に限定した．年次の幅や作品数が限られている条件下で，一定の性能を発揮できたことは TF-IDF によるベクトル化が有効であったと考えられる．特に RFR では予測誤差が小さく，実際の初出年に近い予測ができた．その一方で SVR は他のモデルと比較して予測性能が劣る結果となり，モデルの選択によって差が生じることが示された．交差検証なしでは，いずれのモデルにお

いても誤差が小さく，特に RFR と LR では MAE が 1.11 と非常に優れた性能を得られた．また，先行研究(金[1])では SVR と RFR が優れていたが，本研究では RFR と LR が良好な結果を示し，これは特徴量選択の違いがモデル性能に影響を与えたと考えられる．

表 1. 芥川作品の年次推定の結果(交差検証あり)

モデル	予測の年次誤差 （5回の平均）			
	MAE	MedAE	RMSE	R^2
LR	1.502	1.155	2.061	0.695
SVR	2.282	1.920	2.849	0.448
RFR	1.448	0.954	2.146	0.684

表 2. 芥川作品の年次推定の結果(交差検証なし)

モデル	予測の年次誤差			
	MAE	MedAE	RMSE	R^2
LR	1.111	1.139	1.287	0.886
SVR	2.089	1.711	2.513	0.566
RFR	1.116	0.730	1.725	0.787

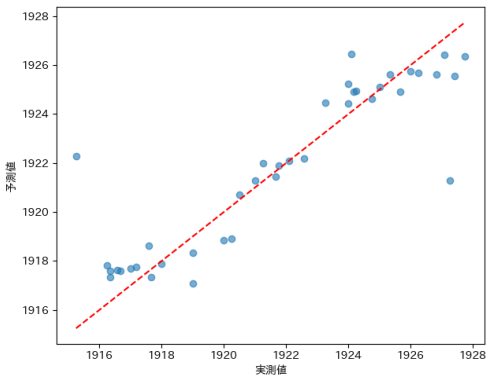

図 1. RFR における実測値と予測値の散布図

5 分析 2:青空文庫作品の初出年次推定

5.1 目的

分析する対象を青空文庫全体にすることで，データセットの規模(対象とする年代幅，作者のバラエティ，作品数，ジャンルなどの規模)を拡大した場合における TF-IDF を用いた手法の有効性を検証することである．また，分析 1 で使用した芥川龍之介の作品を対象としたモデルと比較し，それぞれのモデルにおける性能の違いを明確にし，データセットの規模拡大による影響を明らかにすることを目指す．

5.2 使用したデータセット

データセットは分析 1 で使用した青空文庫収録作品のデータセットを使用した．前処理も同様に MeCab を用いて形態素解析を行い，旧字旧仮名で記述された作品や初出年の情報がない作品，テキストが欠落している作品を除外した．分析 2 では目的変数として「年」の情報のみを使用した．以上のような手順でデータセットを構築し，分析対象とする作品数は 6407 となった．図 3 に分析対象とする作品の年次ごとの分布を示す．

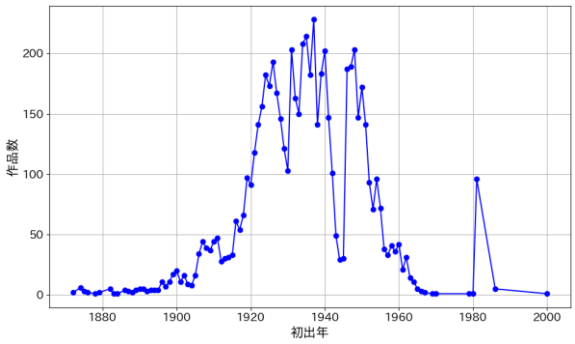

図 3. 初出年ごとの作品数

5.3 実験手順

分析 1 と同様に TF-IDF ベクトル化を用いてテキストをベクトル表現に変換した．この際，最大次元数の選択は分析 1 と同様に，768 の次元数に設定した．分析モデルも LR, SVR, RFR の 3 つのモデルを使用し，データを 5 分割交差検証法に基づき，訓練用 80%とテスト用 20%に分割して学習と評価を行い，交差検証なしの方法でも同様の検証をした．各モデルにおいて，MAE, MedAE, RMSE, R^2 の 4 つの指標を用いて性能を評価した．

5.4 結果と考察

表 3 の結果より，LR では MAE が 8.015 年, MedAE が 6.116 年, RMSE が 10.87 年と，一定の予測性能を示したが，全体的な誤差は分析 1 よりも大きい結果となった．SVR は MAE が 8.426 年, RMSE が 12.19 年と他のモデルよりも高い誤差を示した．一方で RFR ではすべての指標で他のモデルを上回る性能を示し，年次推定において最も優れた結果を示した．また，表 4 の交差検証なしの結果でも，すべてのモデルで交差検証ありとほぼ同程度の性能を示し，交

差検証の有無による性能差はほとんど見られなかった．分析 2 で用いたデータセットのサイズが大きく，多様なデータが含まれていることによって 5 分割時にも全体利用時にも類似した内容構成となったと考えられる．

青空文庫全体を対象とした場合でも，全てのモデルである程度の年次予測が可能であったが，芥川作品のみを対象とした場合と比較すると，全モデルにおいて誤差が大きくなる傾向がみられた．また，青空文庫全体で分析対象のデータに偏りが存在し,特定の年次や作家にデータが集中していることも予測性能に影響したと考えられる．モデルの個々の性能比較では青空文庫全体を対象とした場合でも芥川作品のみを対象とした場合と同様に RFR が最も高い予測性能を示した．

表 3. 青空文庫作品の年次推定の結果 (交差検証あり)

モデル	予測の年次誤差 (5回の平均)			
	MAE	MedAE	RMSE	R^2
LR	8.015	6.116	10.87	0.499
SVR	8.426	5.664	12.19	0.371
RFR	7.588	5.494	10.58	0.526

表 4. 青空文庫作品の年次推定の結果 (交差検証なし)

モデル	予測の年次誤差			
	MAE	MedAE	RMSE	R^2
LR	8.043	6.075	11.25	0.487
SVR	8.628	5.513	12.88	0.326
RFR	7.402	5.110	10.77	0.530

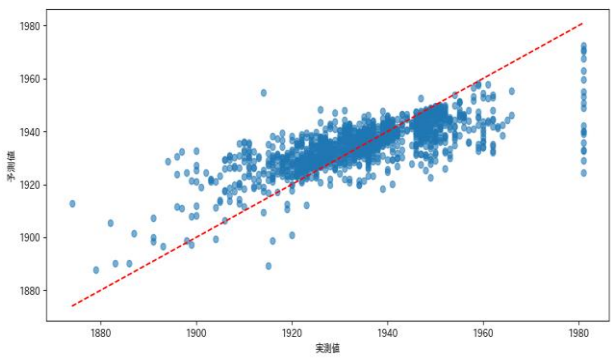

図 5. RFR における実測値と予測値の散布図

6 結論

本研究では，初めに芥川龍之介の作品を対象に初出年の回帰分析を行い, TF-IDF によるベクトル化を用いた複数のモデルで予測性能を比較した．その結果，全てのモデルでかなりの正確さで予測は可能であり，特に RFR モデルで実際の年次に近い予測を実現した．一方で SVR では他のモデルに比べて誤差が大きく，予測性能にばらつきがみられた．

次に青空文庫全体を対象とした場合，全モデルで誤差が芥川作品のみを対象とした場合よりも大きくなる傾向がみられた．分析対象データに偏りがあり，特定の年代や作家にデータが集中していることが予測性能に影響を与えたと考えられる．しかし，全体として初出年の推定は一定の精度で可能であり，特に RFR では高い予測性能を示した．

これらの結果から，芥川龍之介および青空文庫全体の分析おいて, TF-IDF を用いた特徴量設計が初出年の推定に有効であることを示唆された．特に，RFR は多様なデータに対しても安定した性能を示し，年次推定のモデルとして有望であるといえる．RFR は先行研究(金[1])においても優れた成績を示しており，本研究の結果はこれと整合するものである．

7 課題

本研究の課題としてデータの偏りにより，特定の年次や作家，作品ジャンル等に予測性能が影響を受ける可能性がある点が挙げられる．モデルがこれらの偏りに歪められ，一般化性能が低下する懸念がある．次に，特徴量として単語ベースの情報に限定したため，文章全体の構造や執筆スタイルといった文体的要素を十分に反映できてい点が挙げられる．この課題を解決するためには, BERT などの事前学習言語モデルや Word2Vec などの単語分散表現を活用し，単語間の意味的な関連や文脈情報をとらえる手法を導入することが有効であると考えられる．今後は，これらの手法を活用し，モデルの予測精度をさらに向上させる取り組みが求められる．

8 参考文献

[1] 金明哲. 文章の執筆時期の推定—芥川龍之介の作品を例として—. 行動計量学, 2009, 36.2: 89-103.
[2] S. Boldsen and F. Wahlberg. “Survey and Reproduction of Computational Approaches to Dating of Historical Texts,” in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 145–156.
[3] 川崎義史. 永田亮. スペイン語古文書の年次・地点推定のための最適な手法の探求. 言語処理学会 第 28 回年次大会発表論文集, 2022, 1606-1611.
[4] 小林雄一郎. 天笠美咲. 鈴木崇史. 語彙指標を用いた流行歌の歌詞の通時的分析. じんもんこん 2015 論文集, 2015, 2015: 23-30.
[5] 李広微. リコウビ. 近代以降の日本小説の文体変化に関する計量的研究. 2022.
[6]Globis University. 2023. Aozorabunko Clean [データセット]. Hugging Face Hub. https//huggingface.co/datasets/globis-university/aozorabunko-clean. 参照日：2024 年 10 月 28 日