

日本語文学作品における文字 n -gram 特徴量を用いた著者推定の基礎検討

岩崎 優樹¹ 池田 孝利¹

概要：本研究では、青空文庫で公開されている日本語文学作品を対象に、文字 n -gram 特徴量を用いた著者推定の基礎検討を行う。夏目漱石、芥川龍之介、太宰治、森鷗外の4作家、各8作品を対象とし、各作品を1000文字単位の断片に分割した。文字 n -gram ($n = 2, 3, 4$) の TF-IDF 特徴量を抽出し、ロジスティック回帰による分類を行った。評価では、断片ランダム分割と作品単位分割の2条件を比較した。その結果、断片ランダム分割では Accuracy 0.975, Macro-F1 0.973 であったのに対し、作品単位分割では Accuracy 0.810, Macro-F1 0.756 であった。また、作品単位分割では作家ごとの分類性能に差が見られ、太宰治は高い F1 値を示した一方、芥川龍之介は夏目漱石および森鷗外への誤分類が見られた。これらの結果から、文字 n -gram 特徴量は日本語文学作品の著者推定において有効なベースラインとなりうる一方、同一作品から切り出された断片が学習データとテストデータの両方に含まれる条件では、作品固有語や題材語の影響により性能が高く見積もられる可能性が示唆された。したがって、日本語文学作品の著者推定では、断片ランダム分割だけでなく、作品単位分割による評価も併用することが重要である。

A Preliminary Study on Authorship Attribution in Japanese Literary Works Using Character n -gram Features

Abstract: This study presents a preliminary investigation of authorship attribution for Japanese literary works using character n -gram features. We target works by four authors from Aozora Bunko: Natsume Soseki, Akutagawa Ryunosuke, Dazai Osamu, and Mori Ogai. Each work is divided into 1,000-character segments, from which character n -gram features with $n = 2, 3, 4$ are extracted using TF-IDF and classified by logistic regression. We compare two evaluation settings: a random segment split and a work-level split. The random segment split achieves an accuracy of 0.975 and a macro-F1 score of 0.973, whereas the work-level split achieves an accuracy of 0.810 and a macro-F1 score of 0.756. Under the work-level split, classification performance differs across authors: Dazai is classified with high F1, whereas Akutagawa is often misclassified as Soseki or Ogai. These results suggest that character n -gram features can serve as a useful baseline for authorship attribution in Japanese literary works. At the same time, evaluation settings in which segments from the same work can appear in both training and test data may overestimate performance due to work-specific vocabulary and content-related cues. Therefore, authorship attribution for Japanese literary works should be evaluated not only with random segment splits but also with work-level splits.

1. はじめに

文学作品の文体を計量的に分析するスタイロメトリーは、著者推定や作品分析に応用されてきた。著者推定では、文書から抽出した語彙、文字列、品詞、句読点、文長などの特徴量を用いて、未知文書の著者を推定する。日本語文学作品を対象とする場合、旧字旧仮名、文語体・口語体の差、作品の題材、作家ごとの作品数や作品長の偏りなどが評価

に影響する可能性がある。

近年は BERT などの深層学習モデルを用いた著者推定も検討されている。一方で、BERT 系モデルの性能を評価するためには、まず単純で再現可能なベースラインを固定することが重要である。特に文学作品を一定長の断片に分割して分類する場合、同一作品から切り出した断片が学習データとテストデータの両方に含まれると、著者の文体ではなく、作品固有の語彙や内容に基づいて分類できてしまう可能性がある。そのため、文学作品の著者推定では、特徴量や分類器だけでなく、データ分割方法を明確にした評

¹ 大和大学情報学部
Faculty of Informatics, Yamato University

価が必要である。

本研究では、青空文庫の日本語文学作品を対象に、文字 n -gram 特徴量と TF-IDF、ロジスティック回帰を用いた著者推定ベースラインを構築する。本稿では BERT 系モデルまでは扱わず、断片ランダム分割と作品単位分割の 2 条件における文字 n -gram ベースラインの分類性能と誤分類傾向を確認する。本研究の目的は、日本語文学作品の著者推定において、文字 n -gram 特徴量がどの程度有効な基礎的手法となるかを確認するとともに、今後の BERT 系モデルや文体特徴量との比較に向けた評価設計上の論点を整理することである。

本研究の主な貢献は次の通りである。

- 青空文庫の 4 作家、各 8 作品を対象とした著者推定ベースラインを構築した。
- 断片ランダム分割と作品単位分割を比較し、分割条件が分類性能に与える影響を確認した。
- 作家ごとの分類性能と混同行列を確認し、文字 n -gram 特徴量による誤分類傾向を整理した。

2. 関連研究

著者推定およびスタイロメトリーでは、文字 n -gram、単語 n -gram、品詞、句読点、文長などの文体特徴量が広く利用されてきた。文字 n -gram は、形態素解析に依存せずに表層的な文字列の出現傾向を捉えられるため、多様な言語や文書種別に適用しやすい特徴量である。日本語テキストでは、形態素解析や文節構造、助詞・助動詞の使用傾向など、日本語特有の文法的・表層的特徴も著者推定に利用されてきた。

Kanda らは、日本語文学作品を対象に、BERT 系モデルと特徴量ベースモデルを統合した著者推定手法を検討している [1]。同研究は、日本語文学作品を対象とした著者推定において、BERT 系モデルと従来型の特徴量ベースモデルを組み合わせたことの有効性を示している。Liu and Jin は、日本語文書における文節を用いた文体特徴量による著者推定を扱っている [2]。また、Konuma らは、日本語著者推定において、BERT ファインチューニングと文体特徴量を組み合わせた手法を検討している [3]。

これらの研究は、BERT や文体特徴量を用いた高度な著者推定を扱っている。これに対し、本研究では、今後の BERT 系モデルや文体特徴量との比較に向けた基準として、文字 n -gram 特徴量を用いた単純なベースラインを再現可能な形で構築する。特に、同一作品の断片が学習データとテストデータの両方に含まれうる断片ランダム分割と、含まれない作品単位分割を比較する点に着目する。

3. データセット

本研究では、青空文庫で公開されている日本語文学作品を対象とした。対象作家は、夏目漱石、芥川龍之介、太宰

表 1 対象作家と分割後サンプル数

Table 1 Authors and the number of segmented samples.

作家	作品数	サンプル数
夏目漱石	8	73
芥川龍之介	8	70
太宰治	8	90
森鷗外	8	81

治、森鷗外の 4 名である。各作家について 8 作品を選定し、合計 32 作品を用いた。本文データは青空文庫から取得し、著者名、作品名、URL、取得日、文字数、分割後サンプル数を記録した。なお、本文データそのものは GitHub 上では管理せず、取得情報、前処理条件、実験条件、実験結果のみを記録した。

表 1 に、作家ごとの作品数および分割後サンプル数を示す。分割後サンプル数は作家間で完全には一致しないが、各作品から得るチャンク数の上限を 15 に制限することで、長い作品が過度に影響しないようにした。

対象作品は、各作家について短編から中編程度の作品を中心に選定した。ただし、文学作品の長さや題材は作品ごとに異なるため、単純に作品数をそろえるだけではサンプル数は一致しない。そこで本研究では、1000 文字単位の断片に分割したうえで、各作品の最大チャンク数を 15 に制限した。これにより、特定の長い作品が学習データ全体に占める割合を一定程度抑制した。

4. 手法

4.1 前処理

まず、各作品のテキストファイルから、青空文庫の注記、ルビ、メタデータ、底本情報などを削除した。青空文庫テキストには、入力者や校正者、底本情報、ルビ、注記など、本文以外の情報が含まれるため、これらを除去したうえで本文部分を抽出した。その後、抽出した本文を 1000 文字単位のチャンクに分割した。500 文字未満の末尾断片は除外し、各作品から得るチャンク数の上限を 15 とした。

4.2 特徴量と分類器

特徴量として、文字 n -gram を用いた。本実験では、 $n = 2, 3, 4$ の文字 n -gram を抽出し、TF-IDF により重み付けを行った。文字 n -gram は、形態素解析を行わずに、文字列レベルの表層的な出現傾向を捉える特徴量である。分類器にはロジスティック回帰を用いた。実装では、TF-IDF の最小文書頻度を 2 とし、ロジスティック回帰の最大反復回数を 1000 とした。

4.3 評価方法

本研究では、評価条件として断片ランダム分割と作品単位分割の 2 条件を用いた。断片ランダム分割では、1000 文

表 2 作品単位分割における分類性能

Table 2 Classification performance under the work-level split.

指標	値
Accuracy	0.810
Macro Precision	0.801
Macro Recall	0.768
Macro-F1	0.756
Weighted-F1	0.812

字単位に分割した全チャンクをランダムに学習データとテストデータへ分割した．具体的には，全 314 チャンクを作家ラベルに基づいて層化し，乱数シード 42 で学習データ 75%，テストデータ 25%に分割した．その結果，学習データは 235 件，テストデータは 79 件となった．再現性を確保するため，各チャンクに作品ファイル名とチャンク番号に基づく一意な識別子を付与し，学習データとテストデータへの割当を分割一覧（マニフェスト）として保存した．この条件では，同一作品から切り出した断片が学習データとテストデータの両方に含まれる可能性がある．そのため，著者固有の文体だけでなく，作品固有語や題材語，同一作品に特有の表現が分類性能に影響する可能性がある．

作品単位分割では，各作家 8 作品のうち 6 作品を学習データ，2 作品をテストデータとした．この条件では，同一作品から切り出した断片が学習データとテストデータの両方に含まれないようにした．したがって，作品単位分割は，未知作品に対する著者推定性能を確認するための，より厳しい評価条件である．

評価指標には，Accuracy, Macro-F1, Precision, Recall, 混同行列を用いた．サンプル数に偏りがある場合，Accuracy だけでは多数派クラスの影響を受けるため，本研究では Macro-F1 も併せて確認した．主な実験条件は，1000 文字チャンク，最小チャンク長 500 文字，各作品最大 15 チャンク，文字 n -gram 範囲 $n = 2, 3, 4$ ，TF-IDF，ロジスティック回帰である．

5. 結果

作品単位分割における分類性能を表 2 に示す．学習サンプル数は 256，テストサンプル数は 58，特徴量数は 66903 であった．Accuracy は 0.810，Macro-F1 は 0.756 であり，単純な文字 n -gram 特徴量を用いた場合でも一定の著者推定性能が得られることが分かった．

作家別の分類性能を表 3 に示す．太宰治は F1 が 0.980 と高く，多くの断片が正しく分類された．一方，芥川龍之介は Precision が 1.000 である一方，Recall は 0.500 にとどまった．これは，芥川龍之介として予測された断片は正しく芥川龍之介であったものの，芥川龍之介のテスト断片の半数が他作家へ誤分類されたことを示している．

混同行列を図 1 に示す．芥川龍之介の断片は，夏目漱石

表 3 作家別の分類性能

Table 3 Classification performance for each author.

作家	Precision	Recall	F1	Support
夏目漱石	0.538	0.778	0.636	9
芥川龍之介	1.000	0.500	0.667	12
太宰治	1.000	0.960	0.980	25
森鷗外	0.667	0.833	0.741	12

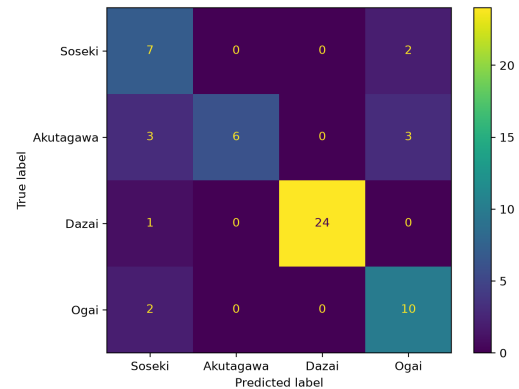


図 1 作品単位分割における混同行列

Fig. 1 Confusion matrix under the work-level split.

表 4 分割条件ごとの分類性能

Table 4 Classification performance under different splitting conditions.

分割条件	Accuracy	Macro-F1
断片ランダム分割	0.975	0.973
作品単位分割	0.810	0.756

および森鷗外にそれぞれ 3 件ずつ誤分類された．また，夏目漱石と森鷗外の間にも一部の誤分類が見られた．太宰治については，25 件中 24 件が正しく分類され，1 件のみ夏目漱石に誤分類された．

表 4 に，断片ランダム分割と作品単位分割の分類性能を示す．断片ランダム分割では，Accuracy 0.975，Macro-F1 0.973 であった．一方，作品単位分割では，Accuracy 0.810，Macro-F1 0.756 であった．断片ランダム分割では，全作家で F1 値が 0.947 以上となり，非常に高い分類性能が得られた．具体的には，夏目漱石と太宰治はすべて正しく分類され，芥川龍之介と森鷗外の断片各 1 件が夏目漱石に誤分類された．

6. 考察

本実験では，文字 n -gram，TF-IDF，ロジスティック回帰を用いた単純なベースラインにより，Accuracy 0.810，Macro-F1 0.756 を得た．この結果から，文字 n -gram 特徴量は，日本語文学作品における著者推定の有効なベースラインとなりうることを示された．文字 n -gram は，単語分割や形態素解析に依存せず，文字列レベルの表層的な特徴を捉えるため，日本語文学作品のように旧字旧仮名や文体

差を含むテキストにも適用しやすい。

断片ランダム分割では、作品単位分割よりも大幅に高い分類性能が得られた。断片ランダム分割では、同一作品から切り出された断片が学習データとテストデータの両方に含まれる可能性がある。そのため、分類器が著者固有の文体だけでなく、作品固有語、題材語、同一作品に特有の言い回しを手がかりとして利用した可能性がある。一方、作品単位分割ではテスト作品が学習時に含まれないため、作品固有語への依存が抑えられ、より厳しい評価条件になる。したがって、日本語文学作品の著者推定では、断片ランダム分割のみで性能を評価するのではなく、作品単位分割での評価を併用することが重要である。

一方で、作家ごとの性能差は大きい。太宰治は高い F1 値を示したが、芥川龍之介は Recall が低く、夏目漱石および森鷗外への誤分類が多かった。この要因として、文語的表現、時代語、作品の題材、語彙の重なりなどが影響した可能性がある。ただし、現時点では、分類器が著者固有の文体を捉えているのか、作品固有語や題材語に依存しているのかは十分に切り分けられていない。

また、本実験では各作品の最大チャンク数を 15 に制限したが、テストサンプル数には作家間で偏りが残っている。例えば、太宰治のテストサンプルは 25 件であり、夏目漱石の 9 件より多い。そのため、今後は作品選定やチャンク数制限の条件を変えた場合の安定性も確認する必要がある。

今後は、誤分類された断片や重みの大きい n -gram を確認し、分類器がどのような文字列特徴に依存しているのかを分析する。さらに、BERT 系モデルや文体特徴量を用いた分類器との比較を行い、文字 n -gram 特徴量との性能差や誤分類傾向の違いを検討する。

7. おわりに

本研究では、青空文庫の日本語文学作品を対象に、文字 n -gram 特徴量を用いた著者推定の基礎検討を行った。夏目漱石、芥川龍之介、太宰治、森鷗外の 4 作家、各 8 作品を対象とし、1000 文字チャンク、TF-IDF、ロジスティック回帰によるベースラインを構築した。

評価では、断片ランダム分割と作品単位分割の 2 条件を比較した。断片ランダム分割では Accuracy 0.975, Macro-F1 0.973 であったのに対し、作品単位分割では Accuracy 0.810, Macro-F1 0.756 であった。この結果から、同一作品から切り出された断片が学習データとテストデータの両方に含まれる条件では、作品固有語や題材語の影響により性能が高く見積もられる可能性があることを確認した。また、作品単位分割では作家ごとの分類性能に差が見られ、太宰治は高い性能を示した一方、芥川龍之介は夏目漱石および森鷗外への誤分類が見られた。

今後は、誤分類された断片や重みの大きい n -gram を確認し、分類器がどのような文字列特徴に依存しているのか

を分析する。さらに、BERT 系モデルや文体特徴量を用いた分類器との比較を行い、文字 n -gram 特徴量との性能差や誤分類傾向の違いを検討する。

参考文献

- [1] Kanda, T., Jin, M. and Zaitzu, W.: Integrated ensemble of BERT- and feature-based models for authorship attribution in Japanese literary works, *Frontiers in Artificial Intelligence*, Vol. 8, Article No. 1624900 (2025). DOI: 10.3389/frai.2025.1624900
- [2] Liu, Y. and Jin, M.: Authorship Attribution Using the Nucleus Bunsetsu as Stylometric Features in Japanese Writings, データ分析の理論と応用 (日本分類学会), Vol. 12, No. 1, pp. 33–46 (2023). DOI: <https://doi.org/10.32146/bdajcs.12.33>
- [3] Konuma, R., Zhang, H., Gao, C. and Wang, J.: Japanese Author Attribution Using BERT Finetuning with Stylometric Features, *Proc. International Conference on Intelligent Multilingual Information Processing (IMLIP 2024)*, Communications in Computer and Information Science, Vol. 2395, Springer, pp. 293–307 (2025). DOI: https://doi.org/10.1007/978-981-96-5123-8_20
- [4] 青空文庫: <https://www.aozora.gr.jp/> (参照 2026-06-18).