# 文体解析を用いた文章の執筆者数推定

黒澤 譲† 矢田 宙生‡ 山名 早人§

†早稲田大学基幹理工学部 〒169-8555 東京都新宿区大久保 3-4-1
‡早稲田大学基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1
§早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

E-mail: †‡§{y.kurosawa, yada_yuki, yamana}@yama.info.waseda.ac.jp

**あらまし** 近年，インターネットの発達により，個人が SNS やブログを利用して情報発信を行う機会が増加している．一方で，信頼性の低い情報の拡散が問題となっている．こうした信頼性の低い情報を判定する一つの方法として,「信頼性の高い文章ほど編集に関わった執筆者の数が多い傾向にある」という点に着目し，文章の執筆者数を推定する手法が提案されている．同手法では，文章を前後に 2 分割する分割点を置き，分割点をずらしながら，分割点前後の文章間での文体類似度が低くなる点を文体変化点して抽出し，執筆者数を推定している．しかし，同手法では，同一執筆者の異なる作品を連結した文章に対して，執筆者数を正しく推定できていない．この問題を解決するために, 本研究では, 文章を複数の区画に分割した上で, 区画のペアに対して文体の類似度を算出することで, 同一執筆者で記述された部分を検出し，執筆者数を推定する手法を提案した．結果，執筆者 1 名の異なる 2 つの作品を連結した文章に対して執筆者数を正しく推定できた確率は，既存手法の 20.0%に対して 99.3%となった.

**キーワード** 情報信頼性，執筆者数推定，文体解析

## 1. はじめに

近年，インターネットの発達により，個人が SNS やブログを用いて情報発信を行う機会が増加した．しかし，フェイクニュースと呼ばれるような虚偽の内容を含む情報が拡散することが社会的問題となっている．最近では新型コロナウイルスが蔓延し，ウイルスに関する誤った情報が広まってしまうと，人々の健康に悪影響を及ぼす可能性がある．Vosoughi ら[1]によると，フェイクニュースは，真実の内容と比較して拡散スピードが速いことが報告されている．以上の背景から，情報の信頼性を自動で判定するシステムの必要性が高まっている.

Wilkinson ら[2]によると，Wikipedia[1]の高品質な記事は他の一般の記事と比較して編集回数と編集者の数が多いことが報告されている． 先行研究[3][4][5]では，この「信頼性の高い文章ほど編集に関わった執筆者の数が多い傾向にある」という点に着目し，文章の執筆者数を信頼性評価の指標として用いた文章の執筆者数推定手法を提案している.

塩浦ら[3]は，品詞の n-gram を用いて文章の特徴量を抽出し，得られた特徴ベクトルに対してクラスタリングを適用することにより執筆者数の推定を行なった．また，渡邉ら[4][5]は，文章を前後に 2 分割し，分割点前後の 2 文章間での類似度を算出し，類似度が閾値以下の場合に文体変化点として検出し，文体変化点の数をもとに執筆者数の推定を行なった．渡邉らの手法では，分割点前後の 2 文章間で類似度計算を行う際，分割点に近い語句の特徴量に高い重みを付与しており，執筆者の境界点を精度よく判定している．しかし，同一執筆者の複数の作品を連結した文章に対しては，執筆者数が 1 人であると正しく推定することができていない．そこで本稿では，その問題を解決するために，文章の特徴量を網羅的に抽出する方法を提案する.

本稿は以下の構成をとる．まず，第 2 節で執筆者数推定の関連研究について述べ，第 3 節で文体解析を用いた執筆者数の推定手法を提案する．第 4 節で提案手法の評価結果を述べ，最後に第 5 節でまとめる.

## 2. 関連研究

本節では，まず文章の特徴量について説明し，次に文章の執筆者数を推定する既存手法について説明する．筆者らの知る限り，文章の執筆者数を推定する手法は塩浦ら[3]の手法と渡邉ら[4][5]の手法のみである.

### 2.1 文章の特徴量

文体解析を行うには，文章の特徴量を抽出する必要がある．文章を構成する単位は，文字，語，文節，文，段落に分けられ，文章構成単位別の特徴量は以下のように分類される[6].

文字：文字種，n-gram，位置
語：品詞，語種，種類，長さ，位置，頻度
文節：係り受け
文：種類，長さ

[1] https://en.wikipedia.org/

段落：長さ，関係性

塩浦ら[3]や渡邉ら[4][5]は，執筆者数推定を行うときに，特徴量として品詞の n-gram を用いている．

### 2.2 クラスタリングを用いた執筆者数推定

塩浦ら[3]は，日本語の文章を対象に，執筆者数の推定を行なった．塩浦らは，まず文章をスライディングウィンドウにより，各ウィンドウ内の特徴量を品詞 n-gram を用いて抽出し，ベクトル化した．次に，得られた特徴ベクトルに対して x-means を用いてクラスタリングを行い，執筆者数を推定した．塩浦らは，品詞 n-gram を用いた特徴ベクトルに対して，重みを$\log(3n)$とすると最も高いクラスタリング精度が得られると報告している．

### 2.3 文体類似度を用いた執筆者数推定

渡邉ら[4][5]は，文体の類似度を算出し，文体の変化を検出することにより，文章の執筆者数を推定した．渡邉らは，まず文章を品詞列に変換し，得られた品詞列を前後に 2 分割した．分割後，分割点近傍の前後をそれぞれベクトル化する．ベクトル化を行うときには，品詞 n-gram の出現頻度を特徴量とし，分割位置に近い程高い重みを付けた．その後，特徴ベクトルを用いて前後の部分の類似度を算出した．以上の処理を，分割点を動かしながら実行し，類似度が変化する点を文体が変わった点として検出し，執筆者数を推定した． その結果，2 人によって記述された文章について，執筆者数の正解率は 81.8%（執筆者数が正解し，かつ文体変化点を正しく推定できた正解率は 65.5%）となった．

また，渡邉ら[5]は，同一執筆者の作品を複数連結することにより 1 人の執筆者で記述された文章を生成し，執筆者数の推定を行なった．同一執筆者の 2 作品を連結した文章では，執筆者を 1 人と推定できた確率は 20%となり，同一執筆者の 3 作品を連結した文章では，執筆者数を 1 人と推定できた確率は 1.5%となった．つまり，同一執筆者の作品を複数連結した文章においては，執筆者数を正しく判定することができていない．

### 2.4 関連研究まとめ

本節では，文章の特徴量について説明し，執筆者数を推定する既存研究の紹介をした．塩浦らはクラスタリングを用いて執筆者数の推定を行い，渡邉らは文章の文体変化を検出することにより執筆者数を推定した．

[2] https://taku910.github.io/mecab/

## 3. 文体解析を用いた執筆者数推定手法の提案

渡邉らの手法[5]では，文章を前後に 2 分割し，分割点近傍に高い重みを付与することにより，前後それぞれの部分に対して特徴量抽出を行い，執筆者が変わる点を精度よく判定している．一方，同一執筆者の異なる作品を連結した文章の場合，作品の境界点を執筆者の境界点として判定してしまうという問題を持つ．この問題を解決するために，本稿では，同一執筆者の複数の作品を連結した文章に対しても，執筆者数を正しく推定することを目指し，分割点近傍の特徴量のみではなく，網羅的に特徴量の比較を行う手法を提案する．

具体的には，文章を複数のブロック（区画と呼ぶ）に分割し，各区画ペアに対して類似度を算出し，類似する連続区画を同一執筆者により記述された部分であるとみなすことで，最終的な執筆者数を推定する．提案手法は，以下の 6 つの処理から構成される．

1. 文章中の単語を品詞に置き換え品詞列を取得する．(特定の固有名詞の影響を抑え，文章の構成や記述の特徴を得るため)（3.1 項）
2. 文章全体の品詞列を複数の区画に分割する．（3.2 項）
3. 各区画に対して品詞 n-gram および skip-n-gram の出現頻度算出と重みづけを行う．（3.3 項）
4. 区画ペアに対して文体の類似度を算出する．（3.4 項)
5. 4 に基づき，同一の執筆者により記述された連続する区画（閉区間と呼ぶ）を検出する．（3.5 項）
6. 5 の結果に基づき，執筆者数を推定する．(3.6 項)

### 3.1 品詞列の取得

本ステップは，渡邉ら[5]の，文章を品詞列に変換する手法と同様である．執筆者数の推定を行うとき，文章中の単語の品詞情報を用いる．これは，文章の内容に依存しない特徴量を抽出するためである．具体的には，対象とする文章に対して形態素解析を行い，単語の品詞情報を 2 階層目まで取得する．形態素解析には MeCab[2]，辞書には NEologd[3]を用いる．ただし，助詞と助動詞に関しては，品詞名ではなく原型を使用する．

表 3.1 取得される品詞情報の例[5]

| 原型 | 1 階層名 | 2 階層目 |
|---|---|---|
| 雨 | 名詞 | 一般 |
| が | 助詞 | 各助詞 |
| 降る | 動詞 | 自立 |
| ． | 記号 | 句点 |
| ， | 記号 | 読点 |

[3] https://github.com/neologd/mecab-ipadic-neologd

### 3.2 区画分割

本ステップでは，3.1 で得られた品詞列に対して，パラメータ$\alpha$を用いて，一つの区画の単語数が$\alpha$以上，かつ各区画の最終単語が句点，かつ各区画の単語数が最小となるように文章を区画に分割する．最終的に，品詞列を$D = \{d_1, d_2, \cdots, d_M\}$のように$M$個の区画に分割できたとする．以下では，$M$個の区画のうち，文章の先頭から数えて$i$番目の区画を区画$d_i$で表す．

### 3.3 特徴量の抽出

区画$d_i$の特徴ベクトル$\boldsymbol{V}_i$の算出方法を以下に示す．特徴量には品詞 skip-2-gram および品詞 3-gram を用いる．全区画での品詞 skip-2-gram，3-gram のうち，出現頻度が 2 以上のものの種類数をそれぞれ$K_2, K_3$とする．$k$番目 $(1 \le k \le K_2)$の品詞 skip-2-gram を$feature_{2,k}$，$k$番目 $(1 \le k \le K_3)$の品詞 3-gram を$feature_{3,k}$とする．区画$d_i$における$feature_{n,k}$の出現頻度を$freq_i(feature_{n,k})$とし，全区画における skip-2-gram，3-gram の延べ出現回数をそれぞれ$sum_2, sum_3$とする．区画$d_i$の特徴ベクトル$\boldsymbol{V}_i$を$\boldsymbol{V}_i = \left(v_i(2,1), \cdots, v_i(2,K_2), v_i(3,1), \cdots, v_i(3,K_3)\right)$とすると，$v_i(n,k)$は式 3.1 で定義される．ここで，$v_i(n,k)$は$k$ 番目の skip-2-gram$(n=2)$および 3-gram$(n=3)$の出現頻度に対する idf を表している．

$$v_i(n,k) = 1 + u_i(n,k) \tag{3.1}$$

ただし，

$$u_i(n,k) = \begin{cases} 0, & if\ freq_i(feature_{n,k}) = 0 \\ \log(4n) \cdot \log\left(\dfrac{sum_n}{freq_i(feature_{n,k})}\right), & otherwise \end{cases} \tag{3.2}$$

### 3.4 文章中の類似度の算出

区画$i$の特徴ベクトル$\boldsymbol{V}_i$と区画$j$の特徴ベクトル$\boldsymbol{V}_j$の類似度の算出には，[5]と同様にコサイン類似度を用いる(式 3.3)．

$$CosSim(i,j) = \frac{\boldsymbol{V}_i \cdot \boldsymbol{V}_j}{|\boldsymbol{V}_i|\,|\boldsymbol{V}_j|} \tag{3.3}$$

### 3.5 同一執筆者により記述された部分の候補推定

3.2 で構築した品詞列の区画$d_i$から区画$d_j$ $(1 \le i < j \le M)$までの連続する区画群を閉区間$[i,j]$と表現する．以下では，同一執筆者により記述された閉区間を求める．具体的には，閾値より大きい類似度を持つ連続した区画を，同一執筆者により記述された閉区間の候補として抽出する．

まず，閉区間の集合$X$を$X \leftarrow \emptyset$と初期化する．全ての$i$，$j$の組$(i,j)$ $(1 \le i < j \le M)$について，以下のⅰ～ⅲを順に実行し，閉区間$[i,j]$が同一の執筆者で記述されているかどうかを判定する．

i. 変数$MinSim_{ij}$を$MinSim_{ij} \leftarrow 1.0$と初期化する．

ii. 全ての$L$，$R$の組$(L,R)$ $(i \le L < R \le j)$ について，$CosSim(L,R)$を計算し，

$$MinSim_{ij} \leftarrow min\left(MinSim_{ij}, CosSim(L,R)\right)$$

と更新する．

iii. パラメータ$th$を用いて，$MinSim_{ij} \ge th$であれば，$X \leftarrow X \cup \{[i,j]\}$と更新する．つまり，閉区間$[i,j]$は同一の執筆者により記述されていると判定し，$[i,j]$を集合$X$に追加する．この時，$X$には，$X$に含まれる閉区間の部分集合となる閉区間も含まれる点に注意する．

閉区間$[i,j]$が同一の執筆者により記述されていると仮定すると，閉区間$[i,j]$内のどの区画のペアに対しても，その類似度が閾値より高くなる可能性が高いと考えられる．したがって，$MinSim_{ij}$が$th$以上となる場合は，閉区間$[i,j]$は同一の執筆者により記述されていると判定する．上記ⅰ~ⅲの処理を実行した後，集合$X$は，同一執筆者により記述されたと判定された閉区間の集合となる．

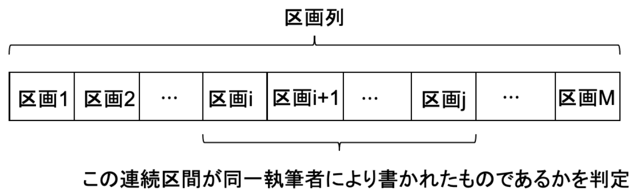


図 3.1 M 区画からなる文章に対する同一執筆者判定

### 3.6 執筆者数の推定

本項では，3.5 までで得られた閉区間の集合をもとに，執筆者数を推定する．具体的には，3.5 までの手順により，「同一の区画$d_i$が複数の閉区間に含まれる場合」を除去することを目的とする．

#### 3.6.1 部分集合となっている閉区間の削除

3.5 までで得られた閉区間の集合$X$について，$X$の要素間で部分集合の関係になっている閉区間を削除する．具体的には，$X = \{x_1, x_2, \cdots, x_{|X|}\}$とすると，$x_i, x_j$ $(i \ne j)$が$x_i \subseteq x_j$ を満たす場合，$x_i$を削除する．

例を図 3.2 に示す．品詞列を$M = 7$個に分割したとし，6 個の閉区間[1,2], [1,3], [2,3], [4,5], [4,7], [5,7]が，それぞれ同一執筆者により記述されたものであると判定されたとする．部分集合の関係になっているものは[1,2] ⊆

[1,3]，[2,3] ⊆ [1,3]，[4,5] ⊆ [4,7]，[5,7] ⊆ [4,7]であるので，[1,2]， [2,3]， [4,5]， [5,7]を削除する．

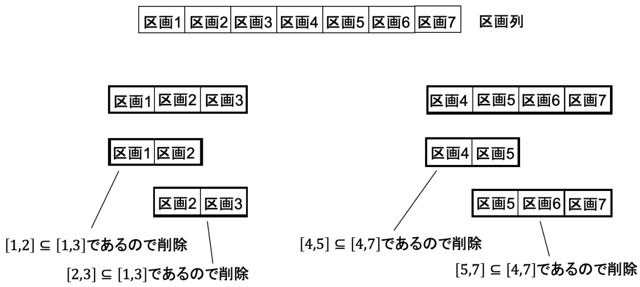

図 3.2 部分集合となっている閉区間の削除

### 3.6.2 閉区間どうしの重なりの削除

3.6.1 までの手順で，閉区間集合$X$における要素間の包含関係は解消することができた．本項では，閉区間集合において，ある区画が重複して複数の閉区間に含まれるような場合を削除することを目的とする．

3.6.1 までの手順で得られた閉区間の集合を$Z = \{z_1, z_2, \cdots, z_{|Z|}\}$とする．このとき，各$z_i$は$z_i = [l_i, r_i]$であるとする（つまり，$z_i$は，品詞列の区画$d_{l_i}$から区画$d_{r_i}$までの連続した区画群である）．ただし，$Z$の各要素は，$l_i$が昇順となるように並んでいるものとする（つまり$l_1 < l_2 < \cdots < l_{|Z|}$であるとする）．

$SimilarityAverage_i$ $(1 \le i \le |Z|)$ を式(3.4)のように定める．

$$SimilarityAverage_i = \frac{\sum_{j=l_i}^{r_i-1} \sum_{k=j+1}^{r_i} CosSim(j,k)}{{}_{r_i-l_i+1}C_2} \tag{3.4}$$

つまり，$SimilarityAverage_i$は，閉区間$z_i = [l_i, r_i]$内の区画どうしの類似度の平均値である．数列$Z = \{z_1, z_2, \cdots, z_{|Z|}\}$を，$SimilarityAverage_i$が昇順になるように並び替えた数列を$Z' = \{z'_1, z'_2, \cdots, z'_{|Z|}\}$とする．

以下の処理ⅰ～ⅲを順に実行することにより，各$z_i$を候補から削除するかどうかを決定し，最終的な執筆者数を推定する．

i　$|Z|$個の変数$delete_1, delete_2, \cdots, delete_{|Z|}$をそれぞれ$False$で初期化する．（つまり，全ての$i$ $(1 \le i \le |Z|)$について$delete_i \leftarrow False$と初期化する）

ii　$i = 1,2, \cdots, |Z|$の順（つまり，$SimilarityAverage$が小さい順）で以下の処理を実行する．

　ii.i　$z'_i = z_j$となる$j$を求める．$1 < j < |Z|$である場合は以下のⅱ.ⅱを実行する．

　ii.ii　$(1 \le k < j \ AND \ delete_k = False)$を満たす$k$のうち，最大のものを$k_1$とする．また，$(j < k \le |Z| \ AND \ delete_k = False)$を満たす$k$のうち，最小のものを$k_2$とする．
　$l_{k_2} \le (r_{k_1} + 1)$である場合，$delete_j \leftarrow True$と更新する．つまり，閉区間$z_j$は，他の閉区間によりカバーされているとみなし，候補から削除する．

iii　$delete_i = False$となる$i$の個数（つまり，候補から削除されずに残った閉区間の個数）を計算し，この値を執筆者数として推定する．

以下では，図 3.3 の例をもとに説明を行う．

- $Z = \{[1,3], [3,5], [4,7], [6,9]\}$であり，区画 3,4,5,6,7 が複数の閉区間に含まれているとする．この時，$SimilarityAverage = \{0.6, 0.5, 0.4, 0.7\}$であると仮定する．$Z$の各閉区間が，同一執筆者で記述された部分の候補である．
- $Z'$は，$SimilarityAverage_i$が昇順となるように$Z$を並び替えたものなので，$Z' = \{[4,7], [3,5], [1,3], [6,9]\}$となる．
- $Z'$の先頭の要素（つまり，$SimilarityAverage$が小さいもの）から，その要素を削除するかどうかを決定していく．
- まず，[4,7]を削除するかどうかを決定する．$Z$において，[4,7]の 1 つ前の項は[3,5]であり，[4,7]の 1 つ後の項は[6,9]である．[4,7]内の区画 4,5 は[3,5]によりカバーされており，また，[4,7]内の区画 6,7 は[6,9]によりカバーされている．したがって，閉区間[4,7]内の全ての区画は，他の閉区間によりカバーされている．したがって，閉区間[4,7]は候補から削除する．
- 次に，[3,5]を削除するかどうかを決定する．$Z$において，[3,5]の 1 つ後の項は[4,7]である．上記の手順において，[4,7]がすでに削除対象となったため，[3,5]は削除しない．
- 次に，[1,3]を削除するかどうかを決定する．[1,3]は，$Z$の先頭の要素であるので，削除しない．
- 最後に，[6,9]を削除するかどうかを決定する．[6,9]は，$Z$の最後の要素であるので，削除しない．
- 以上の処理により，4 つの閉区間のうち，削除対象となったのは[4,7]であり，残りの 3 つの閉区間[1,3], [3,5], [6,9]が候補として残ったので，執筆者数を 3 と推定する．ここで，区画 3 は 2 つの閉区間に属するが，これは区画 3 の途中で筆者が変わったことを示すと考える．

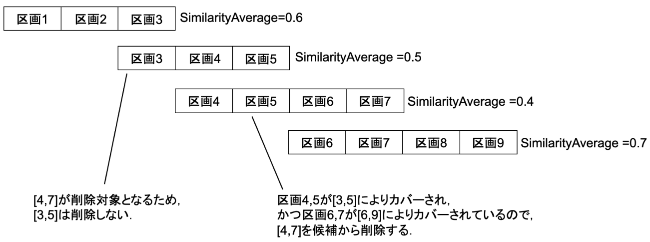


図 3.3 候補区間の削除

## 4. 評価実験

### 4.1 データセット

データセット構築法は渡邉ら[5]が採用した方法と同一である．データセットとして，青空文庫[4]に収録されている小説を用いる．小説の文章は一人の執筆者によって記述されているとみなす．

まず， 2 つのデータセット（パラメータ決定用，テスト用）を作成する．いずれも新字体かつ新仮名遣いで書かれた文章を対象とする．パラメータ決定用データセットでは，青空文庫から無作為に執筆者を 10 名選び，執筆者一人当たり 5 作品，合計 50 作品を無作為に選出し，データセットを作成する．テスト用データセットでは，パラメータ決定用データセットで用いた執筆者と異なる執筆者を無作為に 10 名選び，執筆者一人当たり 10 作品，合計 100 作品を無作為に選出し，データセットを作成する．

以下では，パラメータを決定するためのデータセットをデータセット 1，執筆者数の推定を行うためのデータセットをデータセット 2 と呼ぶ．これらの作成方法を以下に示す．

データセット 1（パラメータ決定用）:

- 執筆者一人あたり 5 作品，異なる 10 人の執筆者の合計 50 作品により構築する．
- 各作品の先頭から 500 単語以上を含み，最初に終わる文までを文群とし，執筆者$i$の作品$j$の文群を$A_{ij}$とする．
- $m$人の執筆者で記述された文章の作成方法：$m$人の執筆者を$B_1, B_2, \cdots, B_m$とし，各執筆者$B_i$に対し，2 つの作品$C_{i1}, C_{i2}$を用いるとする．$m$人の執筆者で記述された文章は，$A_{B_1C_{11}}, A_{B_1C_{12}}, A_{B_2C_{21}}, A_{B_2C_{22}}, \cdots, A_{B_mC_{m1}}, A_{B_mC_{m2}}$をこの順で連結したものである．
- 1 人，2 人，3 人，4 人によって記述された文章を各 100 組，合計 400 組生成した．

データセット 2（執筆者数推定用）:

- 執筆者一人当たり 10 作品，異なる 10 人の執筆者の合計 100 作品により構築した．ただし，ここで用いる執筆者は，データセット 1 で用いたどの執筆者とは異なる．
- $m$人の執筆者で記述された文章の作成方法は，データセット 1 と同様である．
- 1 人，2 人，3 人，4 人によって記述された文章を各 400 組，合計 1600 組を使用した．

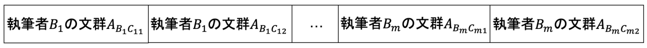


図 4.1 m 人の執筆者で記述された文章

### 4.2 パラメータの決定

提案手法で用いるパラメータの値を決定する．パラメータは，文体の類似度の閾値$th$，分割後の 1 つの区画に含まれる単語数$\alpha$の 2 つである．1 人，2 人，3 人，4 人によって記述された文章を用いて執筆者数の推定を行い，執筆者数を正しく推定する確率が最大となるようなパラメータを求める．パラメータの決定には，データセット 1 を用いる．

推定した文章の執筆者数が実際の文章の執筆者数と一致した場合を正解とし，執筆者数の推定結果が正解となる確率を求めた．この確率が最大となる時のパラメータの組を，グリッドサーチにより求めた．パラメータの値は以下の範囲で変化させた．

$th$:0.05 から 0.95 まで 0.05 ずつ加算
$\alpha$:50 から 250 まで 50 ずつ加算

5 分割交差検証を行い，パラメータの値を決定する．400 組のデータのうち，チューニング用データと検証用データの比を 4:1 とする．5 回の交差検証で，チューニング用データを用いた場合に執筆者数を正しく推定できた確率を最大化するパラメータを求め，各パラメータの平均をとることにより，パラメータの値を決定する．交差検証の結果より，$th$:0.35，$\alpha$:250 とした．

### 4.3 執筆者数の推定

4.2 項で作成したデータセット 2 を用いて，執筆者数の推定を行った．実際の執筆者数と推定した執筆者数が一致した確率（正解率）を表 4.1 に示す．また，実際の執筆者数に対する推定執筆者数の内訳を表 4.2 に示す．

[4] https://www.aozora.gr.jp/

表 4.1 執筆者数の推定確率

| 実際の執筆者数[人] | 執筆者数を正しく推定できた確率（正解率） |
|---|---|
| 1 | 0.993 |
| 2 | 0.318 |
| 3 | 0.305 |
| 4 | 0.193 |
| 1~4 | 0.452 |

表 4.2 実際の執筆者数に対する推定執筆者数

| | 推定した執筆者数[人] | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 実際の執筆者数[人] | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 合計 |
| 1 | **397** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 400 |
| 2 | 238 | **127** | 33 | 2 | 0 | 0 | 0 | 0 | 0 | 400 |
| 3 | 51 | 83 | **122** | 96 | 45 | 3 | 0 | 0 | 0 | 400 |
| 4 | 3 | 15 | 37 | **77** | 113 | 102 | 43 | 9 | 1 | 400 |

1~4 人の執筆者によって記述された文章の正解率は 45.2%となった．なお，1 人の執筆者で記述された文章（2 作品を連結）の正解率は 99.3%となり，従来手法[5]を用いた場合の 20.0%という結果を上回った．

### 4.4 考察

表 4.2 より，実際の執筆者数に対して推定した執筆者数が大きくなるケースがあることがわかる．図 4.2 は，実際の執筆者数が 4 であり，推定した執筆者数が 6 であるケースの一例を示している. 図 4.2 において，[7,8], [12,13], [14,15]のように，連続する 2 個の区画群に対応する閉区間が検出されていることがわかる. 図 4.2 の例以外にも，上記のような閉区間が検出されたケースがいくつか見られた．本手法では$\alpha$（3.2 項）を 250（4.2 項）として使用しており，1 つの区画に 250 単語程度含まれるように文章を分割している. つまり, 2 個の区画群に含まれる単語数は 500 程度である. 一方で，使用したデータセットについて，同一執筆者で記述された部分は 1000 単語程度である（4.1 項）. 以上のことから，本手法は，同一執筆者で記述された箇所を部分的には検出できているが，正確な範囲は検出できていない可能性があると考えられる．したがって，実際の執筆者数よりも推定した執筆者数が大きくなる場合があると考えられる．

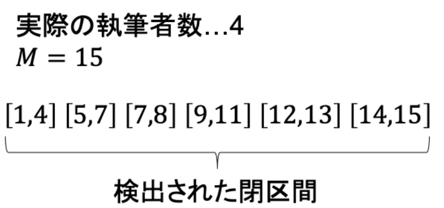


図 4.2 実際の執筆者数が 4 人の文章に対して提案手法を適用した結果の一例

## 5. おわりに

本稿では，文章を複数の区画に分割し，区画のペアに対して類似度の算出することで執筆者数を推定する手法を提案した．品詞 n-gram および skip-n-gram の出現頻度を算出して各部分の特徴を抽出し，n の値に応じた重みづけを行うことで特徴ベクトルを求めた．得られた特徴ベクトルを用いて，各部分のペアに対して文体の類似度を算出し，同一執筆者により記述された部分を検出することで執筆者数を推定した．

評価実験では，同一執筆者の異なる 2 つの文章を連結することで 1 人および複数の執筆者によって記述された文章を生成し，執筆者数を推定した．1~4 人の執筆者により記述された文章に対して提案手法を適用した結果，執筆者数を正しく推定できた正解率は 45.2%となった．また，1 人の執筆者で記述された文章に対して執筆者数を正しく推定できた確率は，既存手法の 20.0%に対して 99.3%となった．しかし，複数の執筆者で記述された文章では執筆者数を正しく推定できた確率が低いため，推定手法を改善し，推定精度を向上させることを検討している．

## 参 考 文 献

[1] S. Vosoughi, D. Roy and S. Aral, “The spread of true and false news online,” *Science*, vol.359, issue 6380, pp.1146-1151, 2018.

[2] D. Wilkinson and B. Huberman, “Cooperation and Quality in Wikipedia,” in *Proc. of WikiSym ’07*, pp.157-164, 2007.

[3] 塩浦尚久, 山名早人, “日本語の文章を対象にした執筆者数推定”, DEIM Forum 2019 論文集, B5-1, 2019.し

[4] 渡邉充博, E. S. Aung, 山名早人, “文体変化と文体類似度を用いた文章の執筆者数推定”, DEIM Forum 2020 論文集, G1-3, 2020.

[5] 渡邉充博, E. S. Aung, 山名早人, “語彙の出現位置と頻度による文体類似度を用いた文章の執筆者数推定”, DEIM Forum 2021 論文集, I21-3, 2021.

[6] 浅石卓真, “テキストの特徴を計量する指標の外

観", 日本図書館学会誌, vol. 63, no. 3, pp. 159-169, 2017.