日本語の文章を対象にした執筆者人数推定

塩浦　尚久† 　山名　早人††,†††

† 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
†† 早稲田大学基幹理工学術院　〒 169–8555 東京都新宿区大久保 3–4–1
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{poche,yamana}@yama.info.waseda.ac.jp

あらまし　近年では，Twitter や Facebook のような SNS が数多く存在し，SNS を通して気軽に文章を投稿できる．しかし，誰でも文章を投稿できるがゆえに，虚偽の情報が広まる可能性がある．虚偽の情報は，政治や経済など社会に影響を与える恐れがある．また，近年の Web 上では，様々な情報が真偽に関わらず公開されており，我々利用者は常にその情報が正しいかどうかを判断する必要がある．本研究では，書籍などの信頼のおける文章は複数人の校閲が入る点にヒントを得，情報の真偽は文章の執筆に携わった人数に関係があるという仮説のもと，日本語の文章を対象に，当該文章が何人で執筆されているのか推定する手法を提案し，推定した文章と情報の真偽に関係があるかどうかの調査することを目指す．具体的には，著者人数は n-gram を用いて文章のベクトル化を行い，クラスタリングを行うことによって推定をする．

キーワード　クラスタリング，執筆者人数推定

1 はじめに

近年では，Twitter[1] や Facebook[2]，ブログといったソーシャルメディアによって，誰でも情報を発信できるようになっている．しかし，真実でない情報が，真実であるかのように発信されてしまう問題がある．例えば，自然災害時に，被災者を混乱させるような情報を発信されるケース [3]が挙げられる．更に，政治目的で虚偽の情報を使用 [4]し，国民の印象を操作する例もある．また，ソーシャルメディアではユーザが発信した情報をシェアすることによって，情報を拡散することができる．この情報のシェアによって，虚偽の情報が拡散されてしまうということも問題となっている．Vosoughi ら [1] は，虚偽の情報は，真実の情報に比べて拡散されるスピードが速いということを報告している．また，ソーシャルメディアを通して誰でも情報を発信できるため，Web 上には情報が多く公開されているが，Web 上に公開されている全ての情報が正確であるとは限らない．なぜなら，情報が古い情報のまま更新されていない，投稿者が誤って発信してしまった，などということが起こり得るからである．このような背景から，近年の Web 上の情報は迂闊に信用できない．したがって，情報を得るために，Web 上の情報が信頼できるかどうかという判断が必要になる．

以上の背景から，Web 上にアップされた情報の真偽を判定する手法が必要である．虚偽のニュースの中でも，SNS に見られる投稿者の勘違いで投稿してしまうというケースは，複数人でチェックを行えば拡散を防ぐことができる．[2] では，複数のユーザによって作成される Web の百科事典である Wikipedia[5] と寄稿者と専任の編集者が作成した百科事典である Encyclopaedia Britannica[6] を比較して，両者の情報の虚偽を調査しており，違いはあまりないという結果が報告されている．これは Wikipedia のように多くの人が文章の執筆に携わることによって，当該の文章がより信頼できる情報になると考えることができる．

本研究では，書籍のような信頼のおける文章は複数人の校閲が入る点にヒントを得，情報の真偽は文章の執筆に携わった人数に関係があると仮定し，執筆者人数を推定する．具体的には，日本語の文章を対象に，文章の作成に携わった人数を文章の文体の変化を用いて推定を行う．これによって，文章から推定した執筆者の人数と，情報の真偽の関係を明らかにすることを目指す．

対象とする日本語の文章は，複数人によって作成された「である調」の文章とする．そして，執筆者人数とは，文章中に現れる異なる執筆者の人数とする．

本稿は以下の構成をとる．まず，第 2 節で関連研究について述べ，第 3 節で提案手法について説明する．そして，第 4 節で実験・評価を行い，最後に 5 節でまとめる．

2 関連研究

筆者らの知る限り，これまでに文章から執筆者人数を推定を行う研究は行われていない．一方，執筆者が一人であるということ前提の元，執筆者を推定する手法が提案されいている．執筆者の推定に関する研究は，いかにして文章から特徴を抽出するかということに焦点を当てている．効果的な特徴量の抽出は，

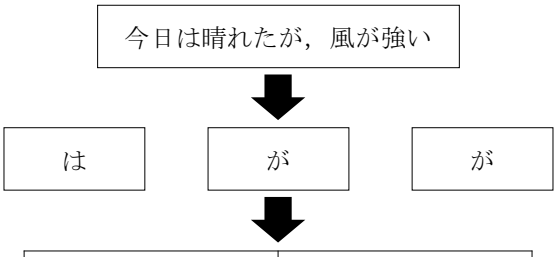
1 : https://twitter.com
2 : https://www.facebook.com
3 : https://www.nikkei.com/article/DGXMZO35227790R10C18A9CC1000
4 : https://mainichi.jp/articles/20161211/k00/00m/030/038000c
5 : https://ja.wikipedia.org/wiki/メインページ
6 : https://www.britannica.com/

対象となる文章によって異なる．本節では，研究対象となった文章と，提案された特徴量を説明する．

2.1 ブログ記事を対象にした研究

中島ら [3] は，ブログを対象に執筆者を推定する研究を行った．中島らは，品詞と助詞の出現パターンを特徴として用いることで，記事の書き方が類似するブログ執筆者を推定する手法を提案している．品詞と助詞の出現パターンを，それぞれ「助詞の n-gram」と「品詞の n-gram」として表現している．

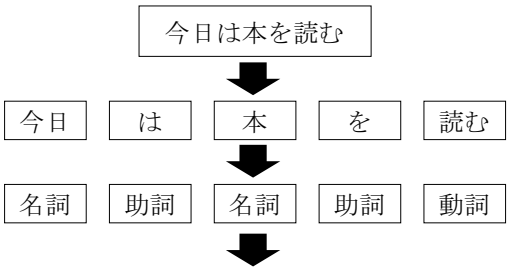
助詞の n-gram は，文章中に出現した助詞を出現順に並べ，n-gram を計算する手法である．図 1 に助詞の n-gram の概要を示す．



パターン(2-gram)	出現数
は，が	1
が，が	1

図 1　助詞の n-gram ([3] 図 1 よりトレース)

品詞の n-gram は，文章に対して形態素解析を行い，品詞名で n-gram の計算を行う手法である．図 2 に品詞の n-gram の概要を示す．



パターン(2-gram)	出現数
名詞，助詞	2
助詞，名詞	1
助詞，動詞	1

図 2　助詞の n-gram ([3] 図 2 よりトレース)

執筆者間の類似度は，式 (1) に示すピアソンの積率相関係数を用いている．

$$corr(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum_{u \in U}(R_{u,i} - \overline{R_i})^2}\sqrt{\sum_{u \in U}(R_{u,j} - \overline{R_j})^2}} \quad (1)$$

ここで，i, j はブログサイト (執筆者) とする．U は品詞と助詞の出現パターンの集合で，$R_{u,i}$ は出現パターン u のブログサイト i における平均出現数，$\overline{R_i}$ はブログサイト i における出現パターンの平均出現数とする．

中島らは，助詞の n-gram の n を 1 から 3 まで，品詞の n-gram の n を 1 から 5 まで変化させて実験を行い，品詞の n-gram に関しては n を大きくすることによって精度が高くなることを報告している．これは，n を大きくすることによって，より複雑な品詞の出現パターンを表現できるようになるためである．

2.2 マイクロブログを対象にした研究

奥野ら [4] は，マイクロブログを対象に執筆者の推定を行った．奥野らは,「品詞タグ・文字混合 n-gram」を提案している．品詞タグ・文字混合 n-gram とは，文章を形態素解析し,「動詞」「接続詞」「記号」「副詞」「形容詞」「感動詞」「未知語」に関しては，文字列をそのまま使用し，それ以外は品詞名に置き換え，n-gram を計算する手法である．品詞タグ・文字混合 n-gram によって，マイクロブログで顕著な未知語を文字列のまま使用することができ，精度を向上した．n-gram は n を 1 から 3 まで変化させ，特徴量として使用した．

2.3 作家の文章及び一般人の作文を対象にした研究

金ら [5] は作家の文章と一般人の作文を対象に研究を行った．金らは，これまでに提案されてきた n-gram 等によって抽出された特徴量は，日本語の構文論のアプローチ [7]がされていないことを指摘し，日本語の文節を用いた特徴量の抽出を行った．

金らは以下に示す複数の文節パターンを提案しており，パターンの頻度によって文章を特徴空間へのマッピングを行った．そして，全てのパターンについて評価実験を行った．

- パターン A: 第 1 層の形態素タグを用いる．
- パターン B: パターン A のうち，助詞と記号はそのままの文字列を用いる．
- パターン C: 第 2 層の形態素タグを用いる．
- パターン D: パターン C のうち，助詞はそのままの文字列を用いる．

ここで，第 1 層の形態素タグとは，品詞名のことを指し，第 2 層の形態素タグとは品詞の細分類のことである．例えば,「文章」という単語の第 1 層の形態素タグは「名詞」であり，第 2 層は「一般」である．それぞれの文節パターンの例を「文章の書き手の識別を行う。」という文章を用いて表 1 に示す．

表 1　文節パターンの例

文節	パターン A	パターン B	パターン C	パターン D
文章の	名詞，助詞	名詞，”の”	一般，連帯化	一般，”の”
書き手の	名詞，助詞	名詞，”の”	一般，連帯化	一般，”の”
識別を	名詞，助詞	名詞，”を”	サ変接続，格助詞	サ変接続，”を”
行う。	動詞，記号	動詞，”。”	自立，一般	自立，”。”

評価実験は以下の 3 種類の文章に対して行われた．

7 : 文節のことを指す

- 作家の文章
- 大学生の作文
- 一般人の作文

結果は，作家の文章ではパターン B，大学生の作文と一般人の作文はパターン C が最も精度が高くなった．

2.4 関連研究まとめ

執筆者の推定で使用されたデータと，執筆者間の差異が大きくなるような特徴量を表 2 にまとめる．既存研究から，全ての種類の文章に対して汎用的な特徴量はなく，対象となる文章に最適化させる必要がある．

表 2 既存研究まとめ

著者	対象文章	特徴量
中島ら [3]	ブログ	品詞と助詞の n-gram
奥野ら [4]	マイクロブログ	品詞タグ・文字混合 n-gram
金ら [5]	作家の文章	表 1 のパターン B
金ら [5]	一般人の作文	表 1 のパターン C

3 執筆者人数の推定

執筆者人数の推定を行う提案手法について述べる．まず，研究の前提条件を以下にまとめる．

- 日本語で書かれた文章である．
- センテンスとは，文，つまり句点で区切られた単位を指す．
- 1 つの文章はセンテンスの集合とする．
- センテンスは 1 名の執筆者によって作成されている．
- 「である」調で作成されている．

提案する手法は，複数のセンテンスからなる一つの文章を入力とし，執筆者の人数を出力する手法である．以下に手順を示す．

（1） 文章をベクトル化する．(3.1 で述べるように，ウィンドウごとにベクトル化を行い，複数のベクトルを得る．以降ベクトル一つ一つをデータと呼ぶ．)
（2） 全データに対して x-means [6] を用いたクラスタリングを行う．
（3） 各データに対して Silhouette 係数 [7] を計算する．
（4） Silhouette 係数が，事前に決定した閾値を下回ったデータを，曖昧なデータとして削除候補とする [8]．
（5） データを出現順 (解析対象の文章におけるセンテンスの出現順) に並べた時，削除候補となったデータの両隣に存在するベクトルのクラスタリング結果が互いに異なる場合 (異なるクラスタに属する場合) は当該データを削除する．
（6） 前の手順で削除されなかった全データに対して x-means を用いたクラスタリングを再度行い，得られたクラスタ数を推定した執筆者人数として出力する．

8：3.2 項で詳細を説明する．

ここで，Silhouette 係数とは，クラスタリング精度を評価する指標である．同じクラスタ同士でお互いに距離が近く，異なるクラスタ同士で距離が遠ければ，良いクラスタであるとして，係数が高いほどよいクラスタリングであることを示す．データ x_i と同じクラスタに属する全てのデータとのユークリッド距離の平均を a^i (凝集度) とし，x_i と最も近い他のクラスタに属する全てのデータとの平均距離を b^i (乖離度) とすると，Silhouette 係数 s^i は (2) 式となる．本手法では，Silhouette 係数に閾値を設け，あるデータが閾値が指定した値より小さければ，曖昧なデータの候補として扱う，

$$s^i = \frac{b^i - a^i}{\max(b^i, a^i)} \tag{2}$$

また，x-means とは k-means [8] を拡張した手法である．k-means と異なり，クラスタ数を事前に指定する必要はなく自動的に決まる．

3.1 n-gram を用いた文章のベクトル化

本項では，文章のベクトル化について述べる．提案する手法では，ベクトル化の方法として，品詞タグと係り受けを利用する手法の 2 種類を提案する．品詞タグを利用する方法は [3] と同様に，「品詞の n-gram」を用いる．文章の形態素解析は MeCab [9]，辞書は mecab-ipadic-NEologd [10] を用いた．

「文節の n-gram」とは，各文章を文節に分解し，文節単位で n-gram を計算する手法である．図 3 に $n = 2$ のときの概要図を示す．ここで，文節の分割は CaboCha [11] を用いた．また，辞書として mecab-ipadic-NEologd を用いた．

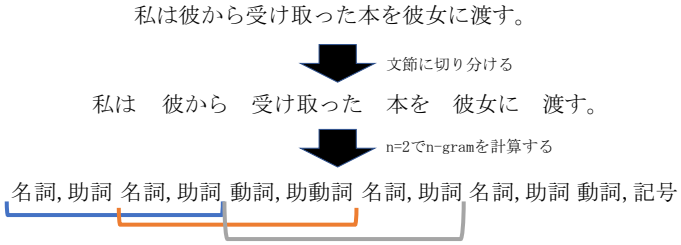


図 3 文節の n-gram の概要図 (n=2 の時)

ベクトル化を行う単位はウィンドウ毎である．図 4 にウィンドウ毎のベクトル化の例を示す．ウィンドウ毎のベクトル化を行うためには，ウィンドウのサイズとストライドを定義する．ウィンドウサイズとは，ベクトル化するときの単語数である．指定したウィンドウサイズを超えて，初めて句点が出現した部分で文章を分割する．ストライドとは，ウィンドウを動かす単位で，センテンス数を指定する．ウィンドウサイズとストライドは任意の値である．図 4 の例では，ウィンドウサイズが 30，ストライドが 1 となっている．

執筆者人数の推定において，文章中で執筆者が変化している

部分が予めわからないため，ウィンドウごとにベクトル化を行う必要がある．また，v_e を n-gram によって分割されたときに得られた形態素の集合 e の出現数 (出現数に関しては図 5 を参考にされたい)，α を非負整数値として，以下に示すように n-gram の n に応じた重み付けを行う．

- v_e
- $\alpha n v_e$
- $\log(\alpha n v_e)$
- $\log(\alpha n) v_e$

本手法では，n の値は 1 から 3 まで変化させて重み付けを行う．

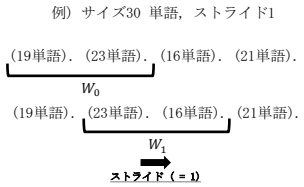
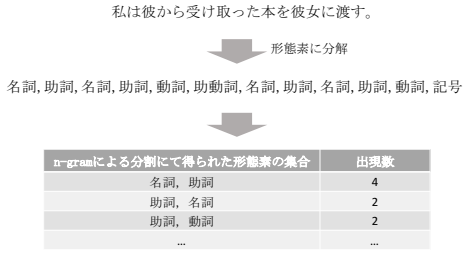


図 4　ベクトル化の例



n-gramによる分割にて得られた形態素の集合	出現数
名詞，助詞	4
助詞，名詞	2
助詞，動詞	2
…	…

図 5　n-gram によるベクトル化における出現数の説明

3.2　曖昧なデータの削除

本項では，曖昧なデータの削除方法について説明する．曖昧なデータとは，1 ウィンドウ内に複数の執筆者が入っているデータである．図 6 で例を示す．ウィンドウ W_2 では，執筆者 a が作成したセンテンスと執筆者 b が作成したセンテンスが混在している．このような場合のウィンドウから得られたベクトルは，クラスタリングの精度を低下させる可能性がある．

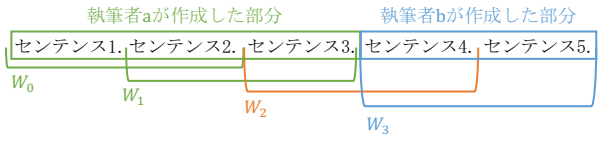


図 6　曖昧なデータが生じる場合の例

ウィンドウに複数の著者が混在しているときの特徴として，該当のウィンドウの前後が互いに異なる執筆者となる．図 6 では，W_2 の 1 つ前のデータ W_1 は執筆者 a，1 つ後のデータ W_3 は執筆者 b の文章となっている．曖昧なデータの特徴より，Silhouette 係数の閾値と，クラスタリングによってデータ列に付与されたラベルの並び順によって判断する．

ここで，データ列とは，図 6 の W_0 から W_3 のように，文章を先頭からベクトル化していく際の順番が付与されたデータである．

クラスタリング後，各ベクトル毎に Silhouette 係数を計算することによって，データのクラスタ内のばらつき (凝縮度) とクラスタ同士の距離の遠さ (乖離度) がわかる．Silhouette 係数に閾値を設け，Silhouette 係数が閾値を下回ることでクラスタから外れたデータを検知する．閾値を下回ったデータは削除候補のデータとする．さらに，削除対象のデータの前後のラベルがお互いに異なっている場合は，削除候補のデータを削除する．6 の例では，W_2 を削除することが目的となる．

4　予 備 実 験

4.1　概　　要

本項では，予備実験に関して説明を行う．予備実験は特徴量の調査を目的として行う．4.2 項でデータセットの説明，4.3 項で予備実験の説明を行う．

4.2　データセット

本研究では，「である調」で作成された日本語の文章を対象にする．理由は，文章の書き方の統一を行うためである．データセットとして本研究室 (早稲田大学情報理工学科または専攻) の中間発表，卒業論文及び修士論文の初稿，またゼミ発表時の予稿を用いた．いずれも概要部分を利用し，合計 20 人分の文章を用いた．

執筆者人数推定の評価をするためには複数の執筆者によって作成された文章が必要である．複数の執筆者によって作成された文章は，1 人の執筆者によって作成された文章をつなげることで擬似的に作成した．

4.3　データの準備

本項では予備実験の説明を行う．概要を図 7 に示す．予備実験では指定人数の文章のベクトル化を行い，クラスタリングを行う．そしてクラスタリングの結果，ある執筆者が書いた文章から得られたベクトル同士が，同じクラスタに含まれるかどうかを調査する．予備実験の目的は，執筆者を分類するための説明能力がある特徴量と，適切な重み付けの調査である．ある特徴量を用いたクラスタリングの精度が高いということは，クラスタリングに使用した特徴量は執筆者によって差異が生じるといえるので，執筆者人数の推定に用いることができる．予備実験の条件は以下のとおりである．

- 執筆者数が事前に決められている．
- 文章の執筆者数は 1 人である．
- 文章の執筆者が既知である．

予備実験で使用したデータは，4.2 項で説明した文章から 10 名分用いて，10 名の中から指定人数分の記事を選んで取得した．

文章のベクトル化について説明する．ベクトル化は，3.1 項で説明した方法とは異なり，ベクトル化の単位をウィンドウ毎では行わずに，重複がないように文章を区切る．予備実験でのベクトル化の例を図 8 に示す．予備実験でのベクトル化では，

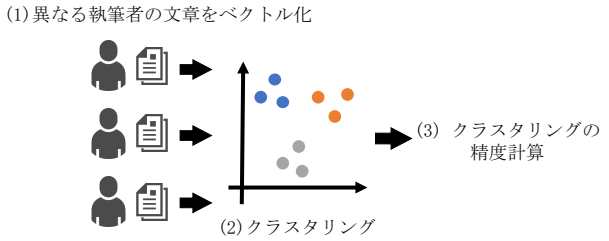


図 7　予備実験の概要

予め指定したウィンドウサイズ (単語数) を超えて初めて句点が出現した部分で文章を区切る．図 8 はウィンドウサイズが 20 の場合である．分割された部分が 1 つのベクトルとなり，図 8 は v_0 から v_2 までの 3 つのベクトルが得られる．予備実験で使用する文章は，複数人で作成された文章を用いず，一人ひとりの文章ごとにベクトル化して行う．よって，1 つの文章の執筆者が 1 名であり，ウィンドウ毎に分割する必要がないため，上記で説明したようなベクトル化を行う．また，ベクトル化する際に，3.1 項で示した重み付けを行う．まずは α を 3 として実験を行う．

例）サイズ 20単語

(20単語)．(23単語)．(16単語)．(21単語)．

v_0　v_1　v_2

図 8　予備実験でのベクトル化の例

予備実験を評価するために，クラスタリングの評価の説明をする．クラスタリングの評価は，執筆者が書いた文章のベクトル同士が同じクラスタに含まれるかどうかで評価を行う．しかし，正解データのラベルとクラスタリングの結果のラベルが一致するとは限らない．図 9 に例を示す．正解データのラベルとクラスタリングの結果のラベルが一致しない問題を解決するために，最もクラスタリングの精度が高くなるようにラベルを付け替えた．図 9 では，2 を 0，0 を 1，1 を 2 と付け替える操作である．クラスタリング後のラベルの付け変えを行うことによって，正確なクラスタリングの評価を行える．

	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
正解のラベル	0	0	0	1	1	1	2	2	2
クラスタリングの結果のラベル	2	2	2	0	0	0	1	1	1

2 → 0, 0 → 1, 1→ 2

	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
付け替え後のラベル	0	0	0	1	1	1	2	2	2

図 9　ラベルの付け替えの例

4.4　重みの決定

予備実験の結果について説明する．予備実験は，10 名の文章の中から，執筆者の人数 k を選び ($_{10}\mathrm{C}_k$ 通り)，全ての選び方に対してクラスタリングを行い，精度の平均値を取得した．予備実験はウィンドウサイズを 100 から 300 に設定して行った．予備実験の結果を図 10 から図 15 に示す．図 10，12 は品詞タグの n-gram の結果で，図 13，15 は係り受けの n-gram の結果である．これらの図は，横軸が執筆者人数であり，縦軸は精度の平均値である．品詞タグの n-gram は $\log(3n)v_e$ が，他の重み付けに比べて最も精度が高い．係り受けの n-gram に関しては $\log(3nv_e)$ が最も精度が高い．

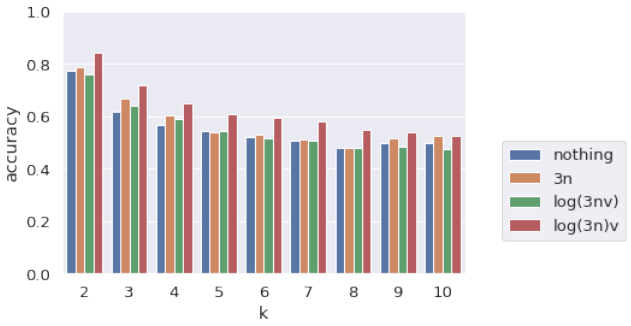


図 10　品詞タグの n-gram，ウィンドウサイズ 100 でのクラスタリング結果

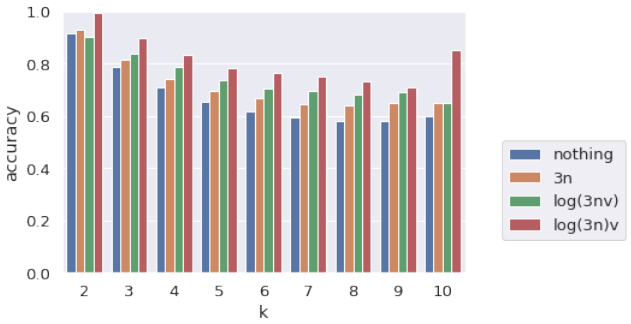


図 11　品詞タグの n-gram，ウィンドウサイズ 200 でのクラスタリング結果

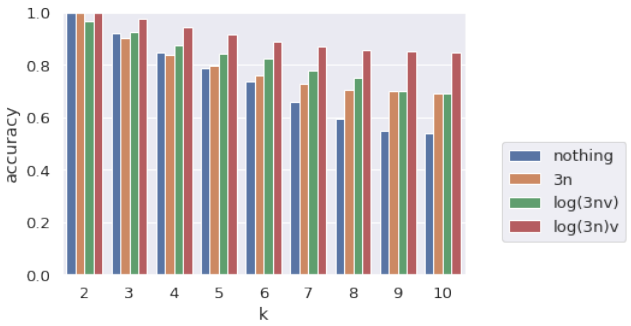


図 12　品詞タグの n-gram，ウィンドウサイズ 300 でのクラスタリング結果

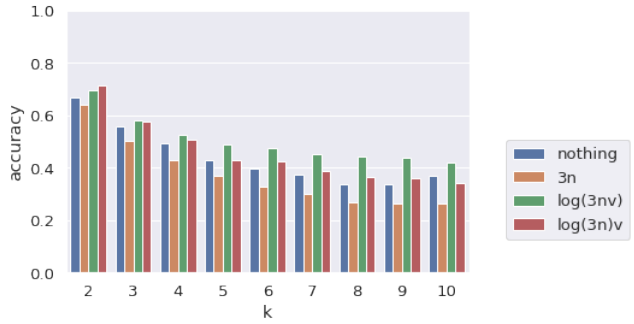


図 13　係り受けの n-gram，ウィンドウサイズ 100 でのクラスタリング結果

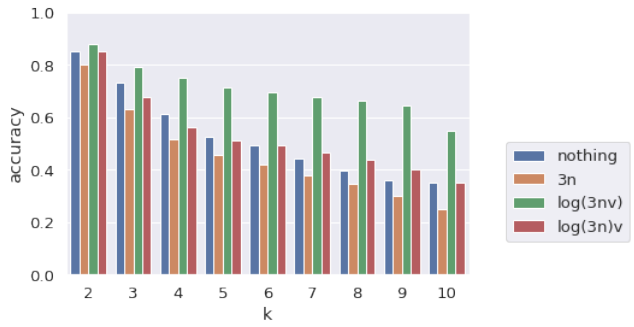


図 14　係り受けの n-gram，ウィンドウサイズ 200 でのクラスタリング結果

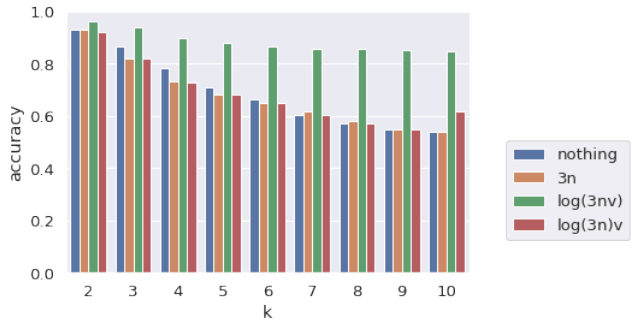


図 15　係り受けの n-gram，ウィンドウサイズ 300 でのクラスタリング結果

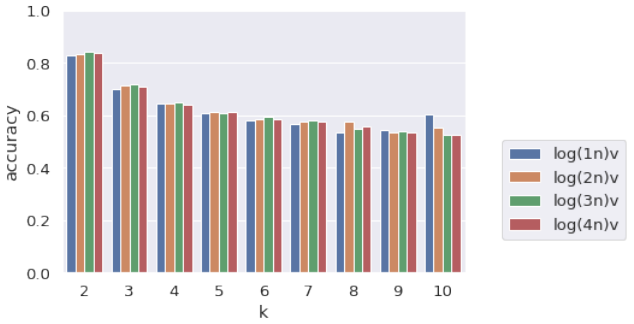


図 16　品詞タグの n-gram，重み付け $\log(\alpha n)v_e$，ウィンドウサイズ 100 でのクラスタリング結果

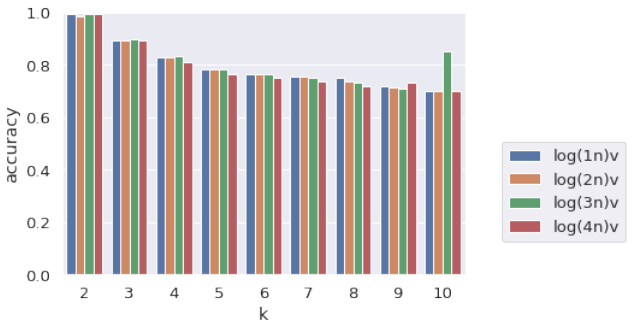


図 17　品詞タグの n-gram，重み付け $\log(\alpha n)v_e$，ウィンドウサイズ 200 でのクラスタリング結果

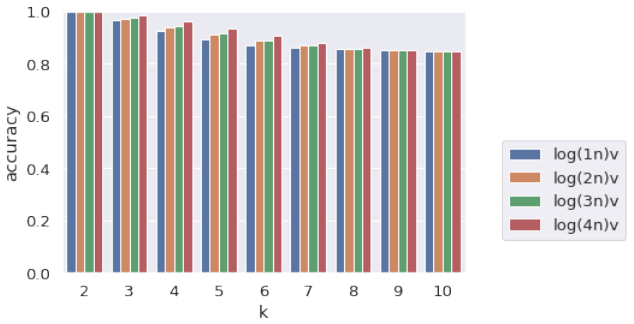


図 18　品詞タグの n-gram，重み付け $\log(\alpha n)v_e$，ウィンドウサイズ 300 でのクラスタリング結果

4.5　重みの係数の決定

次に，重み付けの係数 α の係数を変化させて精度に差異が生じるか実験を行った．図 10 から図 15 の結果を受け，品詞タグの n-gram は $\log(\alpha n)v_e$，係り受けの n-gram は $\log(\alpha n v_e)$ の α を 1 から 4 まで変化させて実験を行った．結果を図 16 から図 21 に示す．結果より，係数 α によって差異がないことがわかった．

4.6　重みの精度の比較

更に，重み付けの係数を $\alpha = 3$ として，品詞タグの n-gram と係り受けの n-gram を組み合わせた場合と，各々を単独で用いた場合の比較を行った．重み付けは，品詞タグの n-gram は $\log(\alpha n)v_e$，係り受けの n-gram は $\log(\alpha n v_e)$ を用いた．結果を図 22 から図 24 に示す．結果から，品詞タグの n-gram のみを用いた結果が最も精度が高いことがわかる．

5　ま　と　め

本研究では，執筆者の人数を推定するための特徴量の調査と，執筆者人数の推定の手法を提案した．特徴量の調査は，品詞タグと係り受けの n-gram と，それぞれの重み付けを決定した．特徴量調査の結果，品詞タグの n-gram で，重み付けは $\log(\alpha n)v_e$ が有効であることを示した．

今後の展望として，今回の調査で有効である特徴量を用いて，実際に執筆者推定を行うことと，執筆者と情報の真偽の関係を明らかにすることが挙げられる．

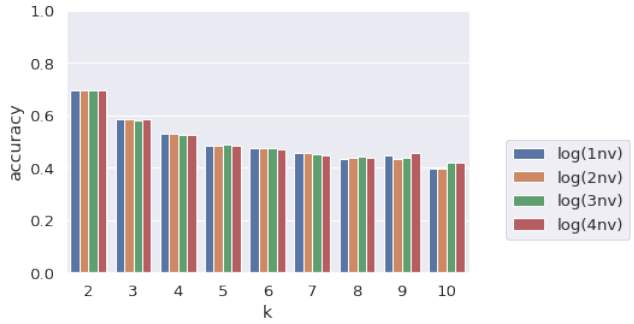


図 19　係り受けタグの n-gram，重み付け $\log(\alpha n v_e)$，ウィンドウサイズ 100 でのクラスタリング結果

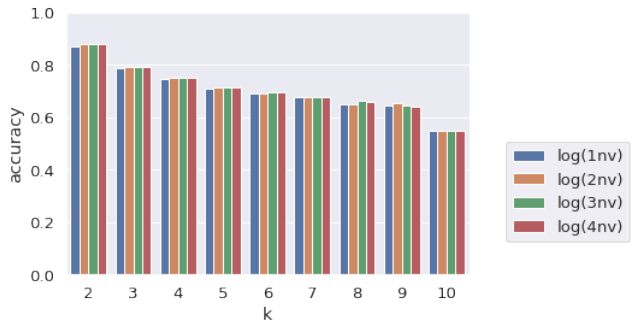


図 20　係り受けタグの n-gram，重み付け $\log(\alpha n v_e)$，ウィンドウサイズ 200 でのクラスタリング結果

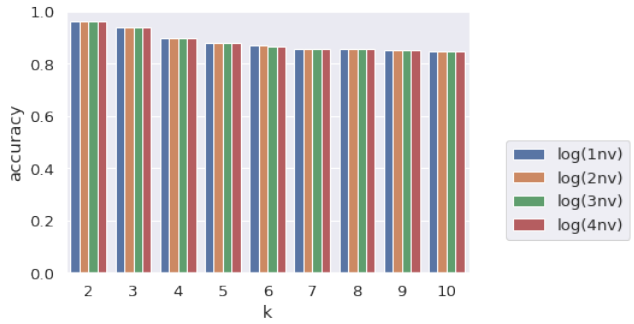


図 21　係り受けタグの n-gram，重み付け $\log(\alpha n v_e)$，ウィンドウサイズ 300 でのクラスタリング結果

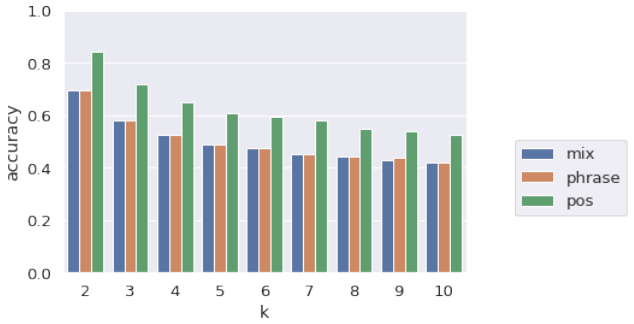


図 22　品詞タグと係り受けを合わせた n-gram，ウィンドウサイズ 100 でのクラスタリング結果

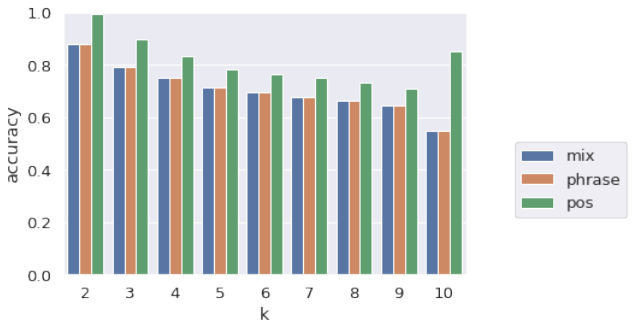


図 23　品詞タグと係り受けを合わせた n-gram，ウィンドウサイズ 200 でのクラスタリング結果

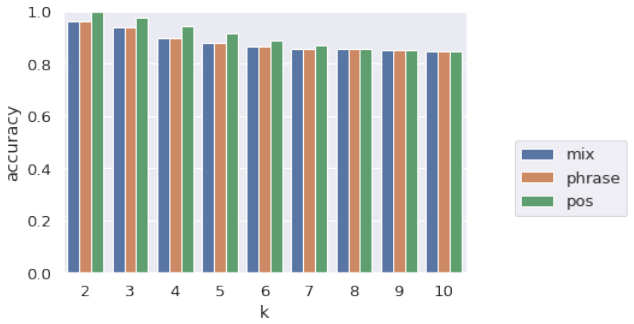


図 24　品詞タグと係り受けを合わせた n-gram，ウィンドウサイズ 300 でのクラスタリング結果

謝　　辞

本研究は科学研究費助成事業 17KT0085 によるものである．

文　　献

[1] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, Vol. 359, No. 6380, pp. 1146–1151, 2018.

[2] Jim Giles. Internet encyclopaedias go head to head. *Nature*, Vol. 438, No. 7070, pp. 900–901, 2005.

[3] 中島泰, 山名早人. 品詞と助詞の出現パターンを用いた類似著者の推定とコミュニティ抽出. 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011), B6-5, 2011.

[4] 奥野峻弥, 浅井洋樹, 山名早人. マイクロブログを対象とした 10,000 人レベルでの著者推定手法の提案. 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014), D7-2, 2014.

[5] 金明哲. 文節パターンに基づいた文章の書き手の識別. 行動計量学, Vol. 40, No. 1, pp. 17–28, 2013.

[6] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, Vol. 1, pp. 727–734, 2000.

[7] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53 – 65, 1987.

[8] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297, Berkeley, Calif., 1967. University of California Press.

[9] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[10] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会発表論文集, pp. 875–878, 2017.
[11] 工藤拓, 松本裕治ほか. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.