

著者推定による文章の特徴解析

國廣 直樹[†] 長谷川 智史[‡] 穴田一[†]

東京都市大学 知識工学部 経営システム工学科[†] 東京都市大学大学院 工学研究科[‡]

1. 研究背景・目的

文章には人それぞれ特徴がある。従来、著者の推定をするための重要な手がかりの 1 つとして筆跡があった。これは人が直接書いた文字の特徴を表し、著者推定以外にも脅迫文解析などに適用されてきた。しかし近年、電子テキストの普及から、筆跡を用いて著者推定をすることはできない文章も増えている。そこで現在、筆跡以外の特徴から文章の著者を推定する研究が盛んに行われている。

これまでの研究[1][3][4]では、n-gram の出現確率や、読点前や文末における各品詞の出現確率など、様々な特徴量を用いて著者推定が試みられている。しかし、未だに著者を推定できていないと言いきれない。それは、著者の特徴を明確に表す特徴量が見つかっていないことが問題であると考えられる。そこで本研究では、著者の特徴を表す特徴量を明確にし、その特徴量を用いて著者を推定することで、その特徴量の有効性を検討することを目的とする。

2. 提案方法

これまでの研究では、文章全体における特徴量を算出し、それを基に著者推定が試みられてきた。しかし、著者の特徴は文章全体ではなく、句点や読点で区切られた文毎に表れるのではないかと考えられる。また、同一文字数の文は文字数が同じであることから、言葉の使い方など、その特徴が出るのではないかと考えた。そこで本研究では、文毎の特徴量を基に著者を推定する方法を提案する。

2.1 文章の加工方法

これまでの研究では、文章の加工をする際に、比較する文章の文字数を統一する方法を用いているものが多く見られる。そのため、その文字数に満たない文章がある場合、その文章を書いた著者の別の文章を繋ぎ合わせるといった加工方法が用いられている。しかし本研究では、このような加工方法は不自然であると考え、文字数の統一を行わず、原文そのままを使用する。また 1 文の文字数を数える際、鍵括弧や括弧などの記号は文字数として数えない。さらに、疑問符や感嘆符で終わる文章は句点で終わる文章として扱う。

2.2 文の定義

本研究では、文章全体ではなく文毎に特徴量を算出する。一般的に、文の定義は句点から次の句点までの文字のことである。この定義を Case①と置く。さらに本研究では、この定義以外に以下の場合も文と定義する。

Case②：読点を句点とし、句点から句点までの間
Case②は句読点の頻度の差が、著者毎に特徴の違いを表すのではないかと考え、提案した。Case①, Case②は、同一文字数の文中でそれぞれ特徴量を算出する。さらに、上記とは別に、句読点が与える影響をより詳しく調べるために、文章を以下のように場合分けをする。

Case③：句点から句点までの間

Case④：句点から読点までの間

Case⑤：読点から句点までの間

Case⑥：読点から読点までの間

上記のように場合分けしたのは、著者によって句読点をつける間隔や、その間に含まれる単語などに特徴が表れると考え、提案した。Case①と Case③の違いは、Case①は単純に句点間の文字をカウントするのに対して、Case③は句点間に読点がないものに対して文字をカウントしている。同様に Case⑥では、読点間に句点が含まれないものに対して文字をカウントしている。そして、それぞれのケースで特徴量を算出し、文章間の非類似度を算出する。

2.3 特徴量

これまでの研究でも、様々な特徴量が提案されてきた。本研究では、既存の特徴量だけでなく、今回新たに考案した特徴量についても著者推定を試みる。本研究で用いた特徴量は、n-gram の出現確率、単語の出現確率、品詞の出現確率、1 文の文字数の出現確率、機能語の出現確率、単語の長さの出現確率、品詞 n-gram の出現確率、単語の長さ n-gram の出現確率の 8 つである。品詞 n-gram とは形態素解析によって得られた品詞に対して、n-gram 法を行ったものである。これは、倒置法などの文章表現方法に著者毎の特徴が表れると考え、提案した。また、単語の長さ n-gram は単語の長さに対して n-gram 法を行ったものである。これは、文字数の長い単語や短い単語の組み合わせをみることで、著者の言葉の使い方に特徴がみられると考え、提案した。本研究では、形態素解析法を行うに際して、奈良先端科学技術大学院大学松本研究室で開発された形態素解析ツールである「茶筌[2]」を用いた。

2.4 非類似度関数

2つの文章間の非類似度を算出する関数は様々あるが,本研究では Tankard[3]が提案した関数を改良し,それを用いる.例えば,品詞の出現確率を特徴量とした時の文章 X , Y の非類似度 $T(X,Y)$ は以下のように表される.

$$T(X,Y) = \frac{A-B}{A} \sum_{a_j} |X(a_j) - Y(a_j)| \quad (1)$$

ここで, A は文章 X , または文章 Y に出現する異なり語数, B は文章 X , Y 双方に出現する共通語数, a_j は文章 X , または文章 Y に出現する品詞, $X(a_j)$, $Y(a_j)$ はそれぞれ文章 X , Y における品詞 a_j の出現確率を表す. この非類似度 $T(X,Y)$ の値が小さければ小さいほど,文章 X , Y は同じ著者によって書かれた文章の可能性が高いことを示している.また結果を比較する為に,松浦ら[4]が考案した次式で表される非類似度関数を用いる.

$$dissim(X,Y) = \frac{1}{card(C)} \sum_{a_j \in C} \left| \log \frac{X(a_j)}{Y(a_j)} \right| \quad (2)$$

ここで, a_j は品詞, C は文章 X , Y 双方に出現する品詞の集合, $card(C)$ は集合 C の要素数, $X(a_j)$, $Y(a_j)$ はそれぞれの文章 X , Y における品詞 a_j の出現確率を示している.この式(2)も式(1)同様,値が小さければ小さいほど文章 X , Y は同じ著者によって書かれた文章の可能性が高いことを示している.式(1)が2つの文章に出現する全ての特徴量 a_j に対して計算を行っているのに対して,式(2)は2つの文章に共通する特徴量 a_j に対してのみ計算を行っている.これらの式(1),式(2)を用いて前述した Case①から Case⑥までの場合それぞれについて,2つの文章間の非類似度を算出する.

2.5 著者推定方法

非類似度を用いて文章の著者を推定する方法について説明する.まず,ある文章を基準文章とする.そして,この基準文章と全ての文章間の非類似度を式(1),式(2)を用いて算出する.次に,得られた非類似度の値を著者毎に足し上げ,その平均を求める.各著者の平均の値を比較し,最も小さい値を取った著者が基準文章を書いた著者と推定する.以上の手順を全ての文章に対して行い,著者推定に成功したその平均成功率で有効性を検討する.本研究では,15 著者,134 作品について解析を行う.文章については青空文庫,BIZ PLUS から引用した.

3. 結果

10 著者,84 作品全てを基準文章とし,2-gram の出現

確率,機能語の出現確率,品詞の出現確率を特徴量とした時の著者推定の平均成功率を以下に示す.

表1 式(1)を用いたときの特徴ごとの平均成功率

	2-gram	機能語	品詞
Case①	22%	67%	18%
Case②	45%	61%	22%
Case③	73%	63%	40%
Case④	87%	74%	25%
Case⑤	75%	71%	43%
Case⑥	68%	64%	20%
Total	80%	70%	39%

2.2 節で定義したそれぞれの場合に対して,2-gram の出現確率,機能語の出現確率,品詞の出現確率を特徴量とした時の著者推定の平均成功率である.Total は Case③から Case⑥で算出された非類似度を足し上げた値に対しての著者推定の平均成功率である.表内の数字が高いほど,著者を正しく推定できたことを示している.

表2 式(2)を用いたときの特徴ごとの平均成功率

	2-gram	機能語	品詞
Case①	19%	48%	21%
Case②	16%	34%	12%
Case③	36%	55%	35%
Case④	39%	43%	23%
Case⑤	46%	40%	30%
Case⑥	46%	51%	40%
Total	53%	49%	40%

2.2 節で定義したそれぞれの場合に対して 2-gram の出現確率,機能語の出現確率,品詞の出現確率を特徴量とした時の著者推定の平均成功率である.Total は Case③から Case⑥で算出された非類似度を足し上げた値に対しての著者推定の平均成功率である.表内の数字が高いほど,著者を正しく推定できたことを示している.

表 1,2 から,この 3 つの特徴量では,n-gram の出現確率が最も著者の特徴を表している.また, Case③から Case⑥の平均成功率が他のケースに比べ高いことから,本研究で提案した文章の場合分けをするこの手法は有用であることが示唆された.その他の結果,考察,今後の課題についての詳細は発表時に述べる.

参考文献

- [1] 金明哲「読点から現代作家のクセを検証する」数理統計 44 121-125(1996)
- [2] 形態素解析ツール「茶筌」
<http://cl.aist-nara.ac.jp/lab/nlt/chasen.html>
- [3] Tankard, J The Literary Detective, BYTE, 11, 2, (1986), 231-189
- [4] 松浦司, 金田康正「近代日本小説家 8 人による文章の n-gram 分布を用いた著者判別」情報処理学会研究報告 53 145-13~8(2000)