4S-1

文章における著者の特徴解析

國廣　直樹†　　長谷川　智史‡　　穴田一‡

東京都市大学　工学部　システム情報工学科†　　東京都市大学大学院　工学研究科‡

1. 研究背景・目的

文章の書き方には人それぞれ特徴がある．その特徴を解析し，著者不明の作品や文献，文章の書き手を識別する研究が行われている．最も特徴が表れるのは文字の筆跡である．これは筆圧，筆勢など様々な要素を組み合わせて，文章の書き手を識別するものである．しかし，直筆で書かれた文章の場合用いることはできるが，ワープロなどで書かれた文章には用いることはできない．現在では，筆跡以外の特徴を用いて，著者を識別するために様々な研究が行われている．

既存の研究では,文章全体における単語の出現頻度など，文章全体から得られる 1 つの特徴を用いて，著者を判別するということが行われている．しかし，私たちは,著者の特徴は文章全体ではなく,同一文字数の文ごとに特徴が変化するのではないかと考えた．そこで本研究では，同一文字数の文ごとの特徴を調べることによって，著者の特徴を抽出することを目的とする．

２．　既存研究

松浦らの研究[1]では，n-gram の出現確率を基に非類似度を定義し，文章の著者を識別することを目的としている.これは，日本語のように単語の切れ目が明確である分かち書きがされていない文章でも,分析可能であることが利点として挙げられる.また，形態素情報や構文情報を必要としないために，対象文章に対して形態素解析などの処理を前もって行う必要がないことも利点である.

この研究では，文章 P，Q 間の非類似度 $dissim(P,Q)$ を次式により定義している.

$$dissim(P,Q)=\frac{1}{card(C)}\sum_{X\in C}\left|\log\frac{P(X)}{Q(X)}\right| \qquad (1)$$

ここで，Xは文字列，C は文章 P，Q 双方に出現する n-gram の集合，$card(C)$ は集合 C の要素数，$P(X)$，$Q(X)$ はそれぞれの文章 P，Q に文字列 X が出現する確率を示している.この文章間の非類似度が小さい値を取るほど，文章 P，Q は同一著者によって書かれたとした.また，比較する文章間の文字数差が大きすぎると，dissim の信頼性が低下するため対象文章の文字数は 3 万字に設定してある．3 万字以上の文章は，先頭から 3 万字を切り出している.3 万字未満の文章には，3 万字に達するまで無作為に同一著者が書いた文章を繋ぎ合わせている.さらに，Tankard らが提案した定義式なども用いて文章間の非類似度を算出し，その値を比較することによって dissim の有効性も検証している.

松浦らは著者を識別するために，まず対象文章から基準文章を 1 つ選択し，この基準文章と他文章間の非類似度を算出した.そして，基準文章を書いた著者の全文章の中で,最も非類似度が大きい文章を選択し,この文章よりも非類似度の値が小さい文章は，基準文章を書いた著者の文章だとした.結果は，n-gram の n が 2 または 3 の時に最も良い値を示した.また，Tankard らが提案した定義式より,松浦らが提案した dissim の方が良い値を示し,dissim の有効性を示したとしている.

この研究では，文字数を統一して解析を行っている.しかし，実際の本や文献，文章の文字数はバラバラである.そこで私たちは,対象文章の文字数がバラバラでも類似度を算出できないかと考えた.さらに，著者の特徴はどの部分に表れるのかを抽出したいと考えた.

3. 提案方法

3.　1　提案概要

私たちは，1 文の文字数は読者が文章を読むリズムに関係するため，著者が意図的に変化させる部分であり，著者の特徴が表れやすいのではないかと考えた．さらに文が同一文字数の時，その文同士は特徴が似ているのではないかと考えた．そこで本研究では，同一文字数の文に着目する．同一文字数の文における n-gram の出現頻度,単語の出現頻度，品詞の出現頻度,品詞のつながりを基に,松浦らの dissim を参考にして新たに非類似度を求める式を定義し，文章間の非類似度を比較した.例えば，n-gram の出現頻度を特徴とした時，文章 X，Y の非類似度 $d(X,Y)$ は以下の式で表される.

$$d(X,Y)=\frac{1}{card(C)}\sum_{i}\sum_{a_j\in C}\left|\log\frac{p(X_i(a_j))}{p(Y_i(a_j))}\right| \qquad (2)$$

ここで X_i，Y_i はそれぞれ文章 X，Y における i 文字の文の集合，a_j は X_i，Y_i 双方に出現する n-gram，C は X_i，Y_i 双方に出現する n-gram の集合，$card(C)$ は集合 C の要素数，$p(X_i(a_j))$，$p(Y_i(a_j))$

Features Analysis of Authors

† Faculty of Engineering, Tokyo City University
‡ Graduate school of Engineering, Tokyo City Univercity

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

はそれぞれ文章 X ， Y における i 文字の文中に a_j が出現する確率を示す．この非類似度 $d(X,Y)$ は小さい値を取るほど，文章 X ， Y は同一著者によって書かれた文章の可能性が高い．また，品詞の出現頻度を特徴とした場合， a_j はそれぞれの文章双方に出現する品詞となる．

3. 2 文の分解方法

日本語の文章は英語などの文章と違い，分かち書きがされていない．したがって，まず文を最小の単位に分解しなければならない．既存の文を分解する方法として，n-gram 法と形態素解析法の 2 つがある．n-gram 法は，文を n 文字ずつの文字列に区切る方法である．このとき，分解された文字列は，単純に文を n 文字ずつ区切っているだけなので，意味を持たない文字列である．一方，形態素解析法は文を最小の単位である単語に区切る方法である．分解された単語は，それぞれ意味を持つ単語であり，品詞判別される．本研究では，両方の方法を用いて解析を行った．なお，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」を用いた．

3. 3 提案手順

文章 X ， Y 間の非類似度を算出するために，以下に本研究における手順を説明する．

STEP①：対象文章における各文の文字数を算出する．ただし，括弧や鍵括弧などの記号は文字数としてカウントしない．また，感嘆符や疑問符で終わる文は句点で終わる文と同じ扱いとした．

STEP②：各文に対して n-gram 法と形態素解析法を用いて，文字列または単語に分解する．このとき n-gram 法の n は 1 から 5 までの値について調べた．また，形態素解析法では，分解された単語の品詞の種類，活用の種類，活用形もそれぞれ調べる．

STEP③：同一文字数の文ごとに，単語の出現頻度などを求め，3.1 節における(2)式を用いて 2 つの文章間の非類似度を算出する．また本研究では，単語の出現頻度の他に，n-gram の出現頻度，品詞の使用率，品詞のつながりを特徴として用いて，文章間の非類似度を算出する．

以上の手順を全ての対象文章に対して行い，算出した文章間の非類似度を比較する．本研究では，8 著者 50 作品について解析を行った．

4. 結果・考察・今後の課題

2-gram の出現頻度を特徴とした時，芥川龍之介の「疑惑」を基準作品とした結果を以下に示す．

表 1 既存の dissim で算出した非類似度の比較

	同一著者の作品群	異なる著者の作品群
平均	7.2443	7.3613
分散	0.02183	0.06739
最大値	7.4851	7.8680
最小値	7.0993	6.7976

上の表は基準文章を書いた著者の他の 5 作品と，異なる著者の作品 44 作品の平均，分散，最大値，最小値である．分散以外の数字が低いほど，基準文章を書いた著者の可能性が高いことを示している．

表 2 提案した $d(X,Y)$ で算出した非類似度の比較

	同一著者の作品群	異なる著者の作品群
平均	0.23788	0.22536
分散	0.00342	0.00252
最大値	0.33861	0.35132
最小値	0.19216	0.16780

上の表は基準文章を書いた著者の他の 5 作品と，異なる著者の作品 44 作品の平均，分散，最大値，最小値である．分散以外の数字が低いほど，基準文章を書いた著者の可能性が高いことを示している．

表 1, 表 2 の結果を見ても分かるように，どちらの非類似度の定義式を用いても，同一著者の作品群と異なる著者の作品群に大きな差はないことが分かる．dissim よりも提案した $d(X,Y)$ を用いて算出した値が極端に小さいのは，同一文字数の文における共通 2-gram の出現頻度が，小さいものばかりになってしまうことが原因だと考えられる．また，基準文章と異なる著者の作品群の中で，基準文章と同一著者の作品群の最大値よりも非類似度が小さい作品は，共に 30 作品ほどあり，全く著者を推定できていないことが分かった．これは，特徴が表れていそうな文の非類似度が計算されていない可能性があることが原因だと考えられる．例えば，一方の文章に i 文字の文が多く出現しても，もう一方の文章に i 文字の文がなければ， i 文字の文において 2-gram の出現頻度を算出しないことを意味している．したがって，提案した定義式をさらに改良する必要性があると考えられる．

また，n-gram の出現頻度には文章の特徴が表れなかったことも考えられる．今後は，文章の特徴を形態素解析により生成された単語の出現頻度，品詞の出現頻度，品詞のつながりにして，さらに検討していく必要があると考えられる．

参考文献

[1]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)

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.