文体測定による筆者識別の一手法

松田　純治・島田　英之*・塩野　　充*・宮垣　嘉也*
岡山理科大学大学院工学研究科修士課程情報工学専攻
*岡山理科大学工学部情報工学科

(1997年10月6日　受理)

1．まえがき

我々が本を読むに際して，人によって読みながら，「好きな文章」，「嫌いな文章」，「読みやすい文章」，「読みにくい文章」などを感じ，それに伴い，「好きな作家」や「嫌いな作家」が生まれてくる。このように文章に対しての印象がなぜ異なるのであろうか。まずは，文章のもととなる言語について少し考えてみる。

言語には，日本語に限らず，文法が存在する。外国語などを学ぶ際にも，文法を教えられる。例えば，英語などでは，文型には5文型あり，それに伴う品詞などのことを授業で説明をする。このように，文法を学ぶことが，外国語を学ぶ上で一種の基本的であり，文章などを書くときもこの文法を意識して書く。このようにして見れば，文章においての文法はたいへん重要な位置をしめているように考えられる。しかし，こと母国語に関して言えば，少々話が違ってくる。母国語の場合，文法の知識を得る前に，すでに話すことができるようになっているので，話すためにあらためて文法を意識することがない。そのため，文章を読む段階になっても，特に意識しない限り，それを気にすることはない。

このようなことから，文章に対しての印象に，文章の根幹とも言える文法というものが大きく関わっているとは考えられない。しかし，もし文章の印象の違いを生むものが何か分かれば，文章の数量化に役立ち，また，書かれた文章の評価も可能になるであろう。

本研究では，文章に対する印象の異なる理由の一つは，文章が「文体」を持つことによるものと考え，著者特有の文体を反映するものとして，句読点の使い方に着目し，それによる識別実験を行った。

2．使用サンプル

本研究において使用したサンプルは，テキストデータである。著者をカテゴリーとして，それぞれ2作品（1万～2万文字）から一定の文字数を1サンプルとし，そのサンプルの約半数を学習サンプルとして辞書パターンの作成に用いた。また，残りを辞書パターンの作成に使用しない，未知サンプルとして用いた。著者と作品は以下の表1に示すとおりである。

表1　使用サンプル

カテゴリー番号	著者	作品名
1	太宰　治	人間失格 斜陽
2	川端康成	伊豆の踊子 雪国
3	夏目漱石	吾輩は猫である こころ
4	司馬遼太郎	関ヶ原 最後の将軍
5	谷崎潤一朗	細雪（上） 細雪（中）

3．識別方法

本研究では，著者が持つ文体は句読点，段落に反映されていると考え，それに着目して識別実験を行った。その方法を以下に示す。

3．1　手法1

1サンプルあたりの「。」，「，」の数を2次元ベクトルと考え単純類似度法を適用する。単純類似度法については以下に説明する。ある特定のカテゴリに属する標準正規データを$g_0(r)$とする。任意に与えられた正規データ$g(r)$がどれほど$g_0(r)$に近いものであるか定量的に測る技法である。$g(r)$のノルム$\|g\|$の値が正確に1に正規化されてない場合を許すことにすれば，適当に選ばれた定数Aに対して$g(r)$が$Ag_0(r)$にどれほど近いかを測定すればよい。そこで$g(r)$と$Ag_0(r)$との違いを両者の差のノルムで評価する立場に立つことにすれば，

$$\|g(r)-Ag_0(r)\|^2=\|g\|^2-2A(g,g_0)+A^2\|g_0\|^2 \tag{1}$$
$$=\|g\|-A\|g_0\|^2+2A\|g\|\|g_0\|=(g,g_0) \tag{2}$$

という恒等関係式が成り立つことになる。この関係から$g(r)=Ag_0(r)$なる関係が成り立つことと，$\|g\|\|g_0\|=(g,g_0)$が成り立つこととは，全くの等価である。そこで特定の$g(r)$に対して，

$$S_s[g(r)]=\frac{(g,g_0)^2}{\|g\|^2\|g_0\|^2} \tag{3}$$

という量を定義して，これを単純類似度とよぶ。また，シュバルツの不等式によれば

$$|(g,g_0)| \leq \|g\|\|g_0\| \tag{4}$$

という不等式が成り立つので，単純類似度は $0 \leq S_s[g(r)] \leq 1$ の範囲の定数値をとり，これを用いた識別方法を単純類似度法という。

3．2　手法2

1サンプルあたりの「。」，「，」の数を2次元ベクトルと考えユークリッド距離を用いて識別実験を行う。ユークリッド距離については以下に説明する。N次元の標準ベクトルMと，未知パターンの特徴ベクトルUのを，$M=(m_1,m_2,\cdots,m_N), U=(u_1,u_2\cdots,u_N)$ とすると，ユークリッド距離 $\|M-U\|$ は

$$\|M-U\| = \sqrt{(m_1-u_1)^2+(m_2-u_2)^2+\cdots+(m_N-u_N)^2} \tag{5}$$

のように表される。

3．3　手法3

「文字」，「。」，「，」を使って波形を生成し，その波形に離散フーリエ変換を適用し，周波数軸データを得る。この得られたデータに対して，単純類似度法を用いて認識を行う。

ここで波形は，「文字」を保持，「。」を±2，「，」を±1としてつくりだす。例えば，『どだい，それは，笑顔ではない。この子は，少しも笑ってはいないのだ。』を波形にすると下の図1のようになる。

4．結果と考察

今回行った実験の結果を以下に順に示す。

4．1　手法1の結果

今回1サンプルあたりの文字数を，500文字単位，700文字単位として，手法1により識

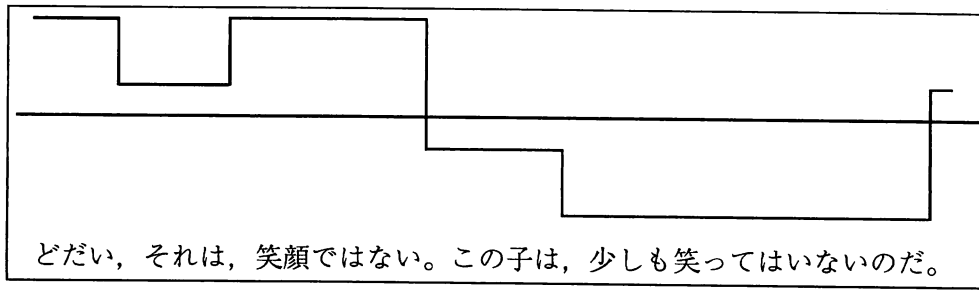

図1　波形の生成

別実験を行った結果を表2に示す。

以上では1サンプルの文字数を一定にしたが，試みに1ページを1サンプルとしたときの結果を下の表3に示す。

表3を見て分かるように全体的に識別率が高く，カテゴリー1とカテゴリー5の識別率も高くなっている。この理由として，本によって1ページあたりの文字数が異なり，それが大きく影響したからと考えられる。しかし，作者によっては，意識的かどうかは不明だが，1行の文字数を調節して，読点を並べたり，避けたり，また同じ語を並べたり，避けたりという部分も見られた。そのような蓄積が多少なりとも結果に反映したと考えられるため，もう少し多くのデータについて重点的に調べる必要がある。

4.2　手法2の結果

表2の結果でカテゴリー1とカテゴリー5を注意深く見ていったときに，識別を失敗した原因は，カテゴリー1ではカテゴリー5に識別してしまうことが多く，カテゴリー5ではカテゴリー1に識別してしまうことが多かったからであることが分かった。カテゴリー1とカテゴリー5の識別率，および互いに間違える確率を以下の表4に示す。

この原因は，両者の文字数単位の句読点の比率が同じためと考えられる。しかし，文字数単位の句読点の数は異なるので，手法2が有効と考え適用した。以下の表5に適用率を

表2　手法1による識別結果

カテゴリー番号	識別率（%）			
	学習サンプル		未知サンプル	
	500/sample	700/sample	500/sample	700/sample
1	52.9	62.5	64.7	70.8
2	50.0	50.0	32.4	50.0
3	70.6	66.7	64.7	66.7
4	67.4	62.5	71.4	70.8
5	70.6	62.5	74.3	79.2

表3　手法1による1ページ単位の識別結果

カテゴリー	学習サンプル	未知サンプル	平均識別率（%）
1	100	95	98
2	70	55	63
3	80	80	80
4	65	50	58
5	100	100	100

示すが，予想通りカテゴリー1とカテゴリー5の識別率が飛躍的に向上した。

4.3　手法3の結果

今回1サンプルあたりの文字数を，500文字単位，700文字単位として，手法3により識別実験を行った結果を表6に示す。

手法1，手法2に比較して識別率は大きく低下しているが，唯一カテゴリー4の識別率

表4　手法1によるカテゴリー1とカテゴリー5の識別結果

カテゴリー番号	識別結果	識別率（%）			
		学習サンプル		未知サンプル	
		500/sample	700/sample	500/sample	700/sample
1	1	52.9	62.5	64.7	70.8
	5	38.2	29.2	33.3	29.2
5	1	23.5	33.3	27.8	21.8
	5	70.6	62.5	74.3	79.2

表5　手法2による識別結果

カテゴリー番号	識別率（%）			
	学習サンプル		未知サンプル	
	500/sample	700/sample	500/sample	700/sample
1	100.0	100.0	97.2	91.7
2	58.3	62.5	41.7	45.8
3	82.4	87.5	68.4	79.2
4	58.8	66.7	75.0	87.5
5	94.1	95.8	100.0	100.0

表6　手法3による識別結果

カテゴリー番号	識別率（%）			
	学習サンプル		未知サンプル	
	500/sample	700/sample	500/sample	700/sample
1	50.0	33.3	41.7	32.0
2	26.5	29.2	28.6	29.2
3	29.4	20.8	22.2	36.0
4	76.5	79.1	80.0	76.0
5	64.7	62.5	69.4	37.5

が，平均的に見て，手法1や手法2のそれよりも高くなっている。このことは，この文字数でもカテゴリー4に特有の周期性が存在することを示していると考えられる。

5．ま と め

今回の実験で極端に識別率が悪かったのはカテゴリー2であったが，この理由として，会話部分が多く，またそれが短いものが多いことが考えられる。また，『雪国』では会話文の「　」の最後に読点を打っているが，『伊豆の踊り子』ではそれがない。この違いが，結果に大きく作用したと考えられる。また2つの作品の発表年代が約10年離れていることも，何らかの影響があるだろう。また，全体的に会話部分が特に短いものを多く含むサンプルや，名詞の羅列を含むサンプルでは，識別率の低下に繋がった。

これを回避するために，今後括弧を特徴として反映するような手法をとることや，1サンプルの文字数を，上記の問題が表面化しない程度にまで増やすなどの対策が必要と考えられる。

また，表6の中で唯一カテゴリー4だけが他の手法の場合より識別率が高くなっている。これはカテゴリー4の著者である司馬遼太郎の文章が，今回実験を行った文字数でも特有の周期性，つまり文章のリズムが認められたからであると考える。以上の実験より，ごく一部の結果からではあるが，文章の特徴が周波数領域の方に顕著に現れる場合が判明した。著者が普通は意識することのない周波数領域の特徴は，筆者識別の有効な特徴となりうるため，今後もより詳細かつ大規模な実験を行う予定である。

参考文献

1）小泉澄之：「フーリエ解析」株式会社朝倉書店，1978.
2）阿居院猛，中嶋正之：「基礎情報工学シリーズ18画像情報処理」森北出版株式会社，1991.
3）飯島泰蔵：「基礎情報工学シリーズ6パターン認識理論」森北出版株式会社，1989.

An Expeiment of Writer Recognition by Measurement of a Style

Junji MATSUDA, Hideyuki SHIMADA*, Mitsuru SHIONO* and Yoshiya MIYAGAKI*

Graduate school of Engineering,

**Department of Information & Computer Engineering,*

Faculty of Engineering,

Okayama University of Science,

Ridai-cho 1-1, Okayama-shi, 700-0005 Japan.

(Received October 6, 1997)

We have an impression when we read a book. For example, we feel that the book "likely", "dislikely", "easy to read" and "hard to read". This fact indicate that we recognizes the character of the novelist unconsciously.

In this paper, we investigated whether or not sentences have some characters. We used three methods for experiment. As a result, using only the distribution of punctuation markes, high recognition rates have obtained about some categories.