エッセイコーパスを用いた著者の生年の推定

岩崎 裕也　　佐藤 理史　　駒谷 和範
名古屋大学 大学院工学研究科 電子情報システム専攻

{yuya_i, ssato, komatani}@nuee.nagoya-u.ac.jp

1　はじめに

テキストには，そのテキストの著者の特徴が様々な形で現れる。それらの特徴を読み取り，性別や年齢などの著者属性を推定する研究が行われている [1, 2, 3, 4]。

本論文では，職業作家が書いたエッセイを対象に，そのテキストの著者の生年を推定する課題を取り上げる。職業作家が書いたエッセイを対象とする理由は，以下の 2 つである。

1. エッセイというジャンルは，小説や論説文などの他のジャンルと比べ，著者の特徴が顕著に現れる。そのため，生年の推定は，相対的に容易であると考えられる。
2. ブログ等の著者と比べ，信頼度が極めて高い生年の情報が得られる。

本研究と同様に，職業作家のエッセイを対象とした著者推定，および，著者の属性推定の研究に，石田ら [3, 5] の研究がある。石田らは，これらの推定に，テキストに出現する文字 bigram を用いた。この研究では，与えられたテキストの著者を，あらかじめ設定した 30 人の著者集合の中から選択するという課題設定で，97.8%の精度を得ている。また，性別推定では，学習テキスト集合に推定対象テキストの著者を含まないという条件下で，最大で 85.6%という結果を得ている。これらの結果は，文字 bigram を用いて著者の特徴を推定することが可能であることを示している。これらの研究結果に基づき，著者の生年の推定を行なう本研究でも，文字 bigram を用いる。

2　推定課題と推定方法

2.1　2 つの推定課題

本研究では，生年推定課題として，次の 2 種類の課題を設定した。

課題 1 与えられたテキスト T の著者の生年が，基準となる年より前か後かを判定する課題（4 節）

課題 2 異なる著者 A，B によって書かれたテキスト T_A，T_B が与えられた時に，著者 A の生年が，著者 B の生年より前か後かを判定する課題（5 節）

これらの課題は，いずれも 2 値分類問題となる。

2.2　素性と素性値

上記の 2 つの課題を解くために，本研究では，SVM (liblinear [6]) を用いて分類器を構成する。

SVM の素性として，有効文字 bigram を使用する。ここで，有効文字とは，ひらがな，カタカナ，JIS 第 1 水準漢字を指し，有効文字 bigram とは，連続する 2 つの有効文字を指す[1]。

課題 1 では，それぞれの有効文字 bigram に対する素性値として，その有効文字 bigram の相対頻度を用いる。テキスト T における有効文字 bigram $\mathbf{x}$ の相対頻度 $\hat{f}(\mathbf{x}, T)$ とは，次式に示すように，有効文字 bigram $\mathbf{x}$ の出現数 $f(\mathbf{x}, T)$ を，すべての有効文字 bigram の出現数の総和で割った値である。

$$v_1(\mathbf{x}, T) = \hat{f}(\mathbf{x}, T) = \frac{f(\mathbf{x}, T)}{\sum_{\mathbf{b} \in B(T)} f(\mathbf{b})} \qquad (1)$$

ここで，$B(T)$ は，テキスト T に出現する，すべての有効文字 bigram の集合（異なり）を表す。

課題 2 では，それぞれの有効文字 bigram に対する素性値として，2 つのテキスト T_A，T_B の有効文字 bigram の相対頻度の差を用いる。すなわち，

$$v_2(\mathbf{x}, T_A, T_B) = \hat{f}(\mathbf{x}, T_A) - \hat{f}(\mathbf{x}, T_B) \qquad (2)$$

2.3　素性選択

有効文字は全部で 3,132 文字あるため，素性の種類は，理論上は，$3,132^2$ 個存在する。しかしながら，実

[1] 文章中に有効文字以外の文字が出現した場合は，その文字を区切りとして，次の文字から再び bigram を抽出する。

際にテキストに出現する素性の種類は，それほど多くはない。

テキスト中に出現する有効文字 bigram の頻度と順位の関係は，語と同様に，Zipf の法則に従う。すなわち，高頻度の bigram は相対的に少なく，低頻度の bigram は相対的に多い。本研究では，信頼度が低い，低頻度の bigram を素性から除外する素性選択を採用する。

この素性選択には，次式で定義される，カバー率 c というパラメータを用いる。

$$c = \frac{\sum_{\mathbf{x} \in U} f(\mathbf{x})}{\sum_{\mathbf{b} \in B(T)} f(\mathbf{b})} \tag{3}$$

ここで，$B(T)$ はテキスト T に出現する全有効文字 bigram の集合，U は素性として使用する有効文字 bigram の集合を表す。すなわち，カバー率 c は，素性として使用する有効文字 bigram の出現数の総和を，全有効文字 bigram の出現数の総数で割った値である。

本研究では，過去の実験等に基づき，カバー率 c が 85%となるように，素性を選んだ。具体的には，頻度の高い順に，有効文字 bigram を，1 つずつ U に追加し，その度に，カバー率を計算する。カバー率が 85%を超えるところまで素性として採用し，さらに，最後に採用した有効文字 bigram と頻度が同じものは，すべて U に含める。そのため，カバー率を 85%と設定しても，実際のカバー率は，この値を少し越えることになる。

3 コーパス

本研究では，以下で説明する，2 つのコーパスを使用した。

3.1 エッセイコーパス

エッセイコーパスは，石田ら [5] によって作成されたコーパスで，職業作家 30 人（男女 15 人ずつ）のエッセイ集から抽出した 900 パッセージ（約 90 万字）から構成されている。著者一人当たりの収録パッセージ数は 30 であり，これらは，異なる 3 冊のエッセイ集から，それぞれ 10 パッセージ（1 パッセージは約 1,000 字）を抽出することにより作成されている。

表 1 に，エッセイコーパスにおける著者の生年と性別の分布を示す。この表より，著者の性別はバランスがとれているが，年代ごとの著者数はバランスがとれておらず（1950 年代が多い），さらに，年代ごとの男女比の偏りも大きい（特に，1930，60 年代）ことが分かる。本研究では，エッセイコーパスを，主として生年推定のための学習データとして使用するが，これらのアンバランスが存在するため，このコーパスは，必ずしも学習データとして最適というわけではない。

表 1: 著者の生年分布（エッセイコーパス）

生年年代	男性	女性	計
1920	3	2	5
1930	5	0	5
1940	2	3	5
1950	4	5	9
1960	1	5	6
合計	15	15	30

表 2: 著者の生年分布（BCCWJ）

生年年代	男性	女性	合計
-1900	14	2	16
1900	4	2	6
1910	15	3	18
1920	50	25	75
1930	30	18	48
1940	41	9	50
1950	16	26	42
1960	7	17	24
合計	177	102	279

3.2 BCCWJ サブコーパス

推定実験のテストデータには，「現代日本語書き言葉均衡コーパス（BCCWJ）」の一部を利用した。

まず，BCCWJ から，日本語十進分類法（NDC）の分類区分が 914（評論，エッセイ，随筆）のサンプル ID を抽出し，さらに，これらの中から，固定長サンプルと可変長サンプルの両方を持つ 338 のサンプル ID を選択した。338 のサンプル ID のうち，59 サンプル ID は，エッセイコーパスと同一著者によるサンプルであり，これらを除外し，残った 279 サンプル ID（著者数は 196 名）を使用した。なお，一つのサンプル ID は，一つのサンプル抽出基準点に対応しており，固定長サンプルはその点から約 1000 字を，可変長サンプルはその点を含む言語的な構造のまとまり（章や節）を，それぞれ抽出することによって作成されている。すなわち，両サンプルは同一テキストから抽出されており，それらの一部は重複している。

表 2 に，279 件のサンプル ID の著者の生年と性別の分布を示す。この表より，1920 年代–50 年代のサンプルが多いことが分かる。また，1940 年代以前は男性のサンプルの方が多く，1950 年代以降では女性のサンプルの方が多い。全体で見ると，男性のサンプル数は女性のサンプル数の約 1.74 倍である。

表 3: 有効文字数の分布

有効文字 bigram 数	サンプル数
-999	2
1,000-1,999	75
2,000-2,999	67
3,000-3,999	44
4,000-4,999	23
5,000-9,999	59
10,000-	9

表 4: 学習データの情報

	のべ数	異なり数
有効文字 bigram 数	701,555	66,827
素性として使用する bigram 数	597,007	12,446
カバー率	85.01%	

可変長サンプルの大きさの分布を，表 3 に示す。有効文字 bigram 数が 10,000 を超えるサンプルもあるが，半分以上のサンプルは，有効文字数 bigram 数が，1000–3000 程度である。

4 生年の二値分類実験

本節では，課題 1，すなわち，与えられたテキスト T の著者の生年が，基準となる年より前か後かを判定する課題に対する実験について述べる。この実験では，学習データにはエッセイコーパスを，テストデータには BCCWJ サブコーパスを用いた。

表 1 に示したように，学習データのエッセイコーパスの著者分布は，1950 年より前に生まれた著者が 15 名，それ以降に生まれた著者が 15 名である。そこで，2 値分類の境界値として 1950 年を採用した。

SVM の学習は，エッセイコーパスの同一著者の同一エッセイ集から得られる 2 パッセージ（約 2,000 字）を，1 つのデータとして用いた。エッセイコーパスは 900 パッセージから構成されているので，学習データの数は 450 個となる。表 4 に，3 節で説明した手法で素性選択を行った際の，有効文字 bigram 数，素性数を示す。実際に使用した素性数は 12,446 種類である。

テストデータ（BCCWJ サブコーパス）279 件に対する実験結果を表 5 に示す。この表に示すように，固定長サンプルに対して 78.2%，可変長サンプルに対して 79.9%の精度が得られた。

表 6 に，表 5 の結果を，10 年毎に集計した結果を示す。この表より，境界値とした 1950 年付近では推定精度が低く，境界値から離れた年代ほど推定精度が高くなる傾向にあることが分かる。

表 7 に，可変長サンプルに含まれる有効文字 bigram 数と精度の関係を示す。この表より，サンプルに含まれる有効文字 bigram 数が増えるほど，推定精度は上昇する傾向があるのが分かる。ほとんどのサンプルにおいて，可変長サンプルの方が固定長サンプルより有効文字 bigram 数が多い。可変長サンプルの推定精度が，固定長サンプルより高いのは，このことが原因と考えられる。

表 5: 生年推定の精度

	固定長		可変長		サン
生年	正解数	精度	正解数	精度	プル
-1949	172	80.8%	175	82.2%	213
1950-	46	70%	48	73%	66
合計	218	78.1%	223	79.9%	279

表 6: 正解の分布

生年	固定長		可変長		サン
年代	正解数	精度	正解数	精度	プル
-1900	14	88%	16	100%	16
1900	6	100%	6	100%	6
1910	18	100%	17	94%	18
1920	70	93%	71	95%	75
1930	33	69%	39	81%	48
1940	31	62%	26	52%	50
1950	28	67%	28	67%	42
1960	18	75%	20	83%	24
合計	218	78.1%	223	79.9%	279

表 8 に，1920-1960 年代生年の著者の精度を男女別に示す。この表より，1940 年代の女性の精度が固定長，可変長の両方おいて，特に精度が低いことがわかる。表 1 を見ると，1940 年代以前が生年の著者は 15 名いるが，女性著者は 5 名しかいない。このことが，推定精度を下げている要因の 1 つと考えられる。

5 著者生年の比較実験

本節では，課題 2，すなわち，異なる著者 A，B によって書かれたテキスト T_A，T_B が与えられた時に，著者 A の生年が，著者 B の生年より前か後かを判定する課題に対する実験について述べる。

まず，テキストの組み合わせの作成方法を説明する。4 節の実験より，生年にある程度の差がないと，識別が困難であると考えられる。そこで，学習データは，T_A，T_B の著者の生年の差が，10 年以上となるような組み合わせを選択した。一方，テストデータは，生年の異なる著者の組み合わせ全てを使用した。

本実験では，エッセイコーパスを学習データ，BCCWJ サブコーパスをテストデータとする実験の他に，BCCWJ サブコーパスを学習データ，エッセイコーパ

表 7: 有効文字 bigram 数と精度の関係

有効文字 bigram 数	正解数	サンプル数	精度
0-4,999	162	211	76.8%
5,000-8,999	48	54	89%
9,000-12,999	13	14	93%

表 8: 男女別精度

	固定長		可変長		サンプル数	
生年年代	男性	女性	男性	女性	男性	女性
1920	94%	92%	98%	88%	50	25
1930	70%	67%	77%	89%	30	18
1940	71%	22%	56%	33%	41	9
1950	63%	69%	63%	69%	16	26
1960	86%	71%	71%	88%	7	17

スをテストデータとする実験も行った。エッセイコーパスは，エッセイ集単位（10,000 字）を 1 つのデータとして扱った。

結果を表 9，10 に示す。これらの表は，2 つのテキストの生年の差と正解数の関係を示している。全体の精度は，BCCWJ サブコーパスをテストデータとした場合は 69.30%，エッセイコーパスをテストデータにした場合は 80.46%となった。前者の設定では，生年差が 30 年以上あるテキスト対に対しては，8 割を超える精度が得られている。しかし，生年差が 10 年以下のテキスト対に対しては，55.58%と精度は非常に低い。

表 9 と表 10 を比較すると，表 10（テストデータ＝エッセイコーパス）の方が精度が高い。この理由は，BCCWJ サブコーパスを学習データとして用いた方が，学習データの量，著者の組み合わせが多かったためだと考えられる。すなわち，より多くの著者を含む学習データを準備することによって，より高い推定精度が得られる可能性がある。

6 おわりに

本論文では，文字 bigram の相対頻度を素性値とした，SVM による著者の生年の推定を行った。テキストの著者の生年が，1950 年より前か，それ以降かを判定する課題において，78.2%の精度が得られた。

また，2 つのテキストの T_A，T_B を与えた時に，どちらの著者の生年が先か，後かを判定する課題においては，エッセイコーパスをテストデータにした時で，80.46%の精度が得られた。

今後の課題として，学習データを増やしての実験，より狭い区切りの年数での推定などが考えられる。

謝辞 本研究では,「現代日本語書き言葉均衡コーパス」の一部を利用した。

表 9: 生年比較結果 (テスト＝ BCCWJ)

生年差	正解数	組み合わせ数	精度
1-10	6,586	11,849	55.58%
11-20	6,221	9,189	67.70%
21-30	5,378	7,207	74.62%
31-40	4,352	5,268	82.61%
41-50	1,551	1,882	82.41%
51-100	2,181	2,541	85.83%
101-120	76	78	97%
合計	26,345	38,014	69.30%

表 10: 生年比較結果 (テスト＝エッセイ)

生年差	正解数	組み合わせ数	精度
1-10	835	1,305	63.98%
11-20	782	1,008	77.58%
21-30	712	756	94.2%
31-45	727	729	99.7%
合計	3,056	3,798	80.46%

参考文献

[1] Jonathan Scheler, Moshe Koppel, Shlomo Argamon and James Pennebakes. *Effects of Age and Gender on Blogging*. 2006 AAAI Spring Symposium Computational Approches to Analyzing Weblogs, pp.191–197, 2006.

[2] Arjun Mukherjee and Bing Liu. *Improving Gender Classification of Blog Authors*. In Proceeding of the 2010 Conference on Empirical Methods in Natural Language Processing. Assosiation for Computational Linguistics, pp.207-217, 2010.

[3] 石田将吾，佐藤理史，駒谷和範. エッセイコーパスを用いたテキストの著者の性別推定. 言語処理学会第 17 年次大会発表論文集, pp.472-475, 2011.

[4] 池田大介，南野朋之，奥村学. *blog* の著者の性別推定. 言語処理学会第 12 回年次大会発表論文集, pp.356-359, 2006.

[5] 石田将吾，佐藤理史，駒谷和範. エッセイコーパスを用いた日本語テキストの著者推定, 情報処理学会 自然言語処理研究会, NL Vol.198 No.6, 2010.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. *LIBLINEAR: A Library for Large Linear Classification*, Journal of Machine Learning Research 9, pp1871-1874, 2008. Software available at `http://www.csie.ntu.edu.tw/~cjlin/liblinear`