エッセイコーパスを用いたテキスト著者の性別推定

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，書き手の特徴が様々な形で現れる。特に，エッセイのようなジャンルのテキストでは，その傾向が顕著である。テキストに現れている様々な特徴を読み取り，著者の人物像を推定することを，著者プロファイリングと呼ぶ [1]。ここでいう人物像とは，著者の性別，年齢，学歴など，その著者に関する属性集合を意味する。

著者プロファイリングは，ブログなどのウェブサイトを利用したマーケティングに応用が可能であると考えられる。たとえば，ブログの著者の性別や年齢が推定できれば，どの層の人間がどのようなことに興味や関心を抱いているか分かるようになる。また，犯罪への科学捜査などへの応用も考えられる。

本論文では，著者プロファイリングの中の性別推定について論じる。まず2節では，関連研究の説明を行う。3節では，実験で用いる手法について説明し，4節では実験で使用する2つのコーパスについて説明する。5節ではエッセイコーパスを用いた実験を，6節ではBCCWJサブコーパスを用いた実験を説明する。

2　関連研究

テキストの著者の性別推定に関する研究のうち，日本語を対象とした研究には，石田ら [3] の研究や，池田ら [4] らの研究がある。

石田らの研究では，推定対象テキストにエッセイを用い，2つの手法の性能を比較している。1つ目は，文字bigramを素性としたSVMを使用する手法で，最大で73.3%の精度を得ている。もう1つの手法は，k-近傍法を用いる方法で，推定対象テキストに対して，そのテキストと類似度が高いテキストを，性別が既知のテキスト集合からk個選び，それらの性別に基づいて最終的な出力を決定する方法である。この方法では，精度83.3%を得ている。

一方，池田らの研究は，ブログテキストの著者性別推定を扱っており，機能語，一人称，形態素などを素性としたSVMを使用して，精度88.9%を得ている。

日本語以外の言語を対象とした性別推定の研究に，英語を対象とした，Jonathanら [1] の研究やArjunら [2] らの研究などがある。Jonathanらの研究は，ブログを対象にしており，品詞，機能語，blog wordなどを素性とし，学習アルゴリズムMulti-Class Real Winnow (MCRW) を用いて80.1 %の精度を得ている。Arjunらの研究もブログを対象に行なった実験であり，SVMの素性として品詞シーケンスを追加し，素性値として既存の手法で使用されている情報利得や相互情報量などを組み合わせて使用している。この研究では，88.56%の精度を得ている。

本研究では，石田らの研究と同様に，エッセイを推定対象テキストとして実験を行う。また，推定手法も石田らの研究に類似した文字bigramを素性としたSVMを使用する。石田らの研究との手法の違いについては，次節で述べる。

3　手法

本研究では，著者の性別推定を2値分類問題として解くことを考える。分類器にはSVM (liblinear) を使用する。SVMの素性には，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第一水準の漢字の計3,132文字を意味する。その他の文字，記号などは無視する。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，その文字の次から再びbigramを抽出する。

3.1　素性値

2節で述べたように，本研究と石田らの研究の手法は，文字bigramを素性としたSVMを用いるという点で類似している。しかし，石田らの研究では，素性値として，文字bigramの生起確率を用いたのに対し，本研究では，素性値として，文字bigramの相対頻度を使用する。

有効文字 bigram ab の生起確率は，テキスト中に文字 a が出現した直後に文字 b が出現する確率 $p(b|a)$ で定義される。この確率は，テキスト中の有効文字 bigram xy の頻度を $f(xy)$ とするとき，次式で定義される。

$$p(b|a) = \frac{f(ab)}{f(a\ )} \tag{1}$$

ここで，　は任意の有効文字を表す。

一方，有効文字 bigram ab の相対頻度は，テキスト中に含まれる有効文字 bigram の頻度の総和に対する，有効文字 bigram ab の頻度の割合，すなわち，

$$\hat{f}(ab) = \frac{f(ab)}{\sum_{xy} f(xy)} \tag{2}$$

として定義する。

本研究では，後者を素性値として採用するが，素性数は最大で 3132^2 となり，そのほとんどの値が 0 になる。そのため，実際に素性として使用するものを，テキスト中に出現する有効文字 bigram のうち，頻度が n 回以上のものに限定する。実験では，このパラメータ n を変えて素性数を変更し，その効果を検証する。

4　コーパス

本節では，実験に使用するコーパスについて説明する。本実験では，エッセイコーパスと BCCWJ サブコーパスの 2 つのコーパスを使用する。

エッセイコーパス　本実験で使用するエッセイコーパスは，石田ら [3] の研究で使用されていたものと同一である。このコーパスは，著者 30 人 (男女 15 人づつ) のテキストを収録している。各著者のテキストは，3 冊のエッセイ集から抽出された 30 パッセージ (1 冊から 10 パッセージづつ) で構成されており，全体で約 30,000 字 (1 パッセージあたり約 1,000 字) である。

BCCWJ サブコーパス　BCCWJ サブコーパスは，「現代日本語書き言葉均衡コーパス (BCCWJ) モニター公開データ (2009 年度版)」から，

1. 日本十進分類法 (NDC) の分類区分が 914(エッセイ) のテキストサンプルで，かつ，
2. 単著で，著者の性別が分かるテキストサンプル

を抽出し，その中から，文字数の近い，男女のテキストを 60 組，計 120 サンプルを選択することによって，作成した。

作成した BCCWJ サブコーパスの概要を表 1 に示す。このコーパスは，以下の点でエッセイコーパスと異なる。

表 1: BCCWJ サブコーパスの概要

文字数	サンプル数
1,000–5,000 字	84
5,000–10,000 字	32
10,000 字以上	4

- 著者数 (120 名) は，エッセイコーパス (30 名) より多い。
- それぞれの著者に対して，テキストサンプルは 1 つしかない。
- テキストサンプルのサイズは，それぞれ異なり，1,000 字程度のものから，10,000 字以上のものもある。なお，10,000 字以上のテキストサンプルは，4 件しかない。

5　エッセイコーパスを用いた実験

5.1　実験 1

実験 1 では，著者 20 名分のテキスト (男性 10 名，女性 10 名) を学習データとして SVM を構成し，残りの著者 10 名分のテキストをテストデータとして使用する。テキストは，同一エッセイから収集された 10 パッセージ (10,000 字) を 1 単位 (インスタンス) として使用する。すなわち，著者 1 名に対し，3 つのデータが存在することになる。

性別推定の精度は，3 分割交差検定により求める。この際，素性としては，学習データ中の総出現回数が n 回以上の有効文字 bigram を用いる。表 2 に，実験で用いた頻度パラメータ n に対する，素性数，および，有効文字 bigram のカバー率を示す。カバー率は，異なり数に対するカバー率と総数に対するカバー率の両者を示した。3 分割交差検定を行なうので，学習データは，3 セット (LS_1, LS_2, LS_3) 存在する。この表では，この 3 セットに対する値の他に，エッセイコーパス全体 (All) に対する値も示した。たとえば，この表で $n = 50$ の All に着目すると，素性数 1716 はエッセイコーパスに出現する有効文字 bigram の異なり数 66772 の 2.88%に過ぎないが，総数 (のべ) に対する割合では 60.8%を占めていることがわかる。

性別推定の実験結果を表 3 に示す。ここで，分類器 C_1, C_2, C_3 は，それぞれ，LS_1, LS_2, LS_3 を学習データとして構成した分類器である。この表より，頻度パラメータ $n = 25$ の時に精度 100%が得られていることが分かる。推定精度は，$n = 25$, 50, 75 では非常に高いが，$n = 1$ や $n = 100$ の場合は劣化する。このことより，高精度の判定結果を得るためには，頻度パラ

表 2: 頻度パラメータ n に対する，素性数と bigram のカバー率 (エッセイコーパス)

		頻度パラメータ n				
		100	75	50	25	1
素性数	LS_1	784	1047	1600	3275	62314
	LS_2	795	1035	1572	3285	60387
	LS_3	781	1038	1588	3257	62190
	All	884	1119	1716	3480	66772
異なり数	LS_1	1.26%	1.68%	2.57%	5.26%	100%
	LS_2	1.32%	1.71%	2.60%	5.44%	100%
	LS_3	1.26%	1.67%	2.55%	5.24%	100%
	All	1.41%	1.88%	2.88%	5.76%	100%
総数	LS_1	48.4%	52.3%	57.9%	67.7%	100%
	LS_2	49.1%	52.7%	58.3%	68.3%	100%
	LS_3	48.4%	52.2%	57.8%	67.6%	100%
	All	51.2%	55.1%	60.8%	70.3%	100%

表 3: 頻度パラメータと推定精度 (実験 1)

	頻度パラメータ n				
	100	75	50	25	1
C_1	28/30	28/30	28/30	30/30	25/30
C_2	15/30	30/30	30/30	30/30	27/30
C_3	28/30	28/30	29/30	30/30	26/30
合計	71/90	86/90	87/90	90/90	78/90
精度	79%	96%	97%	100%	87%

表 4: 入力テキストサイズと精度 (実験 2)

入力サイズ	10,000 字	5,000 字	1,000 字
C_1	30/30	58/60	266/300
C_2	30/30	59/60	256/300
C_3	30/30	56/60	248/300
合計	90/90	173/180	770/900
精度	100%	96.1%	85.6%

表 5: 頻度パラメータ n と推定精度 (実験 3)

	頻度パラメータ n				
	100	75	50	25	1
C_1	87/120	85/120	89/120	87/120	78/120
C_2	75/120	84/120	90/120	92/120	89/120
C_3	87/120	89/120	86/120	90/120	85/120
合計	249/360	258/360	265/360	269/360	252/360
精度	69.2%	71.7%	73.6%	74.7%	70.0%
C_A	94/120	98/120	99/120	95/120	91/120
精度	78.3%	81.7%	82.5%	79.2%	75.8%

メータ n (あるいは，素性数) を適切に設定する必要があることがわかる。この実験からは，有効文字 bigram の総数の 50%から 70%程度をカバーするような頻度パラメータ n を選ぶのがよい，ということが示唆される。

この実験で得られた精度は，石田らの研究 [3] で，学習データに同一著者を含まない条件下で得られた精度 73.3%より，格段に高い。このことから，SVM の素性値としては，文字 bigram の生起確率よりも，文字 bigram の相対頻度の方が，優れていることがわかる。

5.2 実験 2

実験 2 では，実験 1 で構成した分類器 ($n = 25$) を用い，入力テキストサイズを，10,000 字から 5,000 字または 1,000 字に縮小すると，推定精度がどのように変化するかを調べた。なお，入力テキストサイズを縮小すると，著者 1 名に対するテストデータの数 (入力数) が増加する。たとえば，1,000 字の場合は，著者 1 名当たりのテストデータの数は 30 となる。

実験結果を表 4 に示す。この表より，入力テキストのサイズを小さくすると，推定精度が低下することがわかる。5,000 字の場合の精度は 96.1%と高いが，1,000 字の場合は，85.6%に低下する。このことから，今回，我々が採用した方法で，非常に高い精度の性別判定を実現するためには，5000 字程度のテキストが必要であることがわかる。

6 実験 3

エッセイコーパスによる実験の一般性を確認するために，BCCWJ サブコーパスに対する著者の性別推定実験を実施した。この実験では，実験 1 の 3 分割交差検定の際に作成した 3 種類の分類器 (C_1, C_2, C_3) に加え，エッセイコーパス全体を用いて学習した分類器 (C_A) を用い，BCCWJ サブコーパスのそれぞれのサンプルテキストの著者の性別を，正しく推定できるかどうかを調べた。実験結果を表 5 に示す。

この表より，エッセイコーパスの 2/3 を使用して構成した分類器 C_i $(i = 1, 2, 3)$ より，エッセイコーパス全体を学習データとして使用して構成した分類器 C_A の方が性能がよいことがわかる。C_A は，より多くのテキストを学習データとして使用しているので，この

表 6: テキストサイズと精度 (実験 3)

	テキストサイズ		
	10,000 字以上	5,000–10,000 字	1,000–5,000 字
C_1	4/4	24/32	61/84
C_2	3/4	24/32	63/84
C_3	4/4	23/32	59/84
合計	11/12	73/96	181/252
精度	92%	76%	71.8%
C_A	4/4	24/32	71/84
精度	100%	75%	85%

結果は当然である。

注目すべき点は，分類器 C_i $(i = 1, 2, 3)$ と分類器 C_A とでは，性能がピークとなる頻度パラメータ n が異なることである。前者は $n = 25, 50$ の場合が性能が良いが，後者は $n = 50, 75$ の場合が性能が良い。この事実と表 2 の結果を総合すると，頻度パラメータ n よりも，有効文字 bigram の総数に対するカバー率 r の方が，分類性能と強い相関があることが示唆される。すなわち，n をパラメータとして採用するのではなく，実際に使用する有効文字 bigram の (総数に対する) カバー率 r をパラメータとして採用した方が，異なるサイズの学習データを使用して構成した分類器の性能比較に有用であると考えられる。

表 5 の実験結果を，実験 1 の結果 (表 3) と比較すると，推定精度がかなり低下していることがわかる。4 節で述べたように，BCCWJ サブコーパスのテキストサンプルのサイズは，それぞれの著者で異なり，その 70%は 5000 字未満である。そこで，$n = 50$ の場合の結果を，入力テキストのサイズ毎に集計した。その結果を表 6 に示す。分類器 C_i $(i = 1, 2, 3)$ に対する結果は，予想通り，サイズが小さいほど性能が低下していた。しかし，分類器 C_A の結果は，予想とは異なり，5,000–10,000 字のテキストに対する推定精度が最も低かった。この原因は，今後，追求する必要があるが，分母が小さいため，十分に信頼できる数値にはなっていない可能性がある。

表 6 と表 4 で，分類器 C_i $(i = 1, 2, 3)$ に対する結果を比較すると，テキストサイズの条件を揃えても，本実験の判定精度は，実験 1 の判定精度に対して，かなり低い値となっていることがわかる。この事実と，1,000–5,000 字のテキストに対する分類器 C_A の性能 (85%) を総合して判断すると，学習データには，より多くの著者のテキストを含める必要があるということが示唆される。

同じ性別であっても，それぞれの著者によって，テキストに現れる特徴は異なる。著者の性別推定は，それを，男性と女性の 2 グループに自動分類するわけであるが，そのためには，それぞれの性別に対して，より多くの特徴のバリエーションを学習しておくことが望ましい。このバリエーションが乏しいと，多様な著者の性別を正しく推定することは困難である。エッセイコーパスを用いた 3 分割交差検定の実験では，1 つの分類器は，10 名の著者の性別推定しか行なっていない。これに対し，BCCWJ サブコーパスを用いた本実験では，120 名の著者の性別を推定している。推定対象とするテキストの著者数の差が，2 つの実験の推定精度の差として現れたと解釈するのが妥当である。

7 おわりに

本論文では，文字 bigram の相対頻度を素性値として利用した，SVM によるテキストの著者性別推定法を提案し，同一のコーパスを用いた以前の研究に比べ，高い推定精度が得られることを示した。

テキストの著者推定は，対象とするテキストのジャンル，コーパスサイズ，推定対象とするテキストサイズなどによって，問題の難しさが大きく異なるため，他の研究と単純に推定精度を比較することに大きな意味はない。しかしながら，著者数 30 名のエッセイコーパスで著者集合を 3 分割する交差検定において，96–100%の精度が得られたことは，大きな前進である。一方で，120 名の著者に対する性別推定の精度は 82.5%であり，十分に高いとは言えない。今後は，学習データに，より多くの著者を含めることで，推定精度の向上を図っていく予定である。

謝辞 本研究では,「現代日本語書き言葉均衡コーパス」モニター公開データ (2009 年度版) の一部を利用した。

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.