エッセイコーパスを用いたテキストの著者の性別推定

石田 将吾　　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

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

1　はじめに

あるテキストが与えられたとき，そのテキストだけから著者の人物像を推定することは可能であろうか．ここで人物像とは，具体的に性別，年齢，学歴などを指す．このような問題に対する研究は，著者プロファイリング（Author Profiling）と呼ばれる．著者プロファイリングは，マーケティング調査など様々な目的に利用できると考えられている．

本論文では，人物像の属性の一つである性別をテキストから推定する問題を扱う．まず2節で，使用するコーパスについて説明する．3節でk近傍法を用いた性別推定法及び実験，4節で二値分類器を用いた性別推定法及び実験について述べ，5節でまとめる．

2　エッセイコーパス

本稿で述べる性別推定実験では，主にエッセイコーパス [1] を利用した．このコーパスは，職業作家30人（男性15人，女性15人）に対し，エッセイ集（単行本）をそれぞれ3冊選び，そのそれぞれから約1,000字のテキストを10ヶ所，抽出・電子化することによって作成されたコーパスである．以下では，抽出した約1,000字単位のテキストをパッセージと呼ぶ．著者1人当たりの文字数は，約1,000字 × 10パッセージ × 3冊 = 約30,000字であり，コーパス全体では，約900,000字である．

3　k近傍法を用いた性別推定

3.1　方法

ここでは，あるテキストQの著者の性別を推定することを考える．これを実現する一つの方法として，そのテキストと類似したテキストをいくつか見つけ，それらのテキストの著者の性別の多い方を，Qの著者の性別として出力する方法が考えられる．これは，k近傍法を用いて性別推定を実現することに相当する．

具体的な手順は，次のようになる．

1. テキスト集合$\mathbf{T} = \{T_1, T_2, \cdots, T_n\}$を準備する．ここで，$T_i$の著者と性別は既知であるとする．以下では，$\mathbf{T}$を参照テキスト集合，$T_i$を参照テキスト，参照テキストの著者を参照著者と呼ぶ．
2. 任意のテキストと参照テキスト間の類似度を計算する方法を定義する．
3. 著者の性別を推定したいテキストQが与えられる．これを推定対象テキストと呼ぶ．
4. 推定対象テキストQとの類似度が大きい参照テキストを，$\mathbf{T}$からk個選ぶ．
5. 選ばれたテキストの著者の性別の多い方をQの著者の性別として出力する．

この手順で最も重要な部分は，類似度の定義である．本研究では，類似度の計算に，文字bigram言語モデルに基づく尤度を用いる．まず，各参照テキストT_iに対して，文字bigram言語モデルM_iを作成する．テキストQと参照テキストT_iの類似度は，T_iから作成した言語モデルM_iに対するQの尤度$L(M_i|Q)$として定義する．すなわち，

$$\begin{aligned} sim(Q, T_i) &= L(M_i|Q) \\ &= \sum_{x_j x_k \in Q} f(x_j x_k, Q) \log \hat{P}_i(x_k|x_j) \end{aligned}$$

ここで，$f(x_j x_k, Q)$は，テキストQに現れる文字列$x_j x_k$の頻度，$\hat{P}_i(x_k|x_j)$は，テキストT_iにおいて，文字x_kが文字x_jに後続する(補正された)確率を表す．この尤度は，昨年我々が実施した，著者推定の研究で用いた尤度と同一である．詳細は，文献 [1] を参照されたい．

なお，上記の方法は，文献 [1] の著者推定法のある種の拡張とみなすことができる．事実，ある著者集合のそれぞれの著者に対して参照テキストT_iを一つづ

表 1: 実験 1 の結果

	k					
n	1	3	5	7	9	11
1	**84.4**	78.9	78.9	75.6	76.7	77.8
2	**92.2**	84.4	84.4	84.4	82.2	78.9
3	**93.3**	86.7	83.3	87.8	82.2	80.0
4	**95.6**	87.8	85.6	82.2	82.2	83.3
5	**96.7**	86.7	87.8	84.4	82.2	80.0
10	**97.8**	91.1	88.9	85.6	83.3	78.9

表 2: 実験 2 の結果

	k					
n	1	3	5	7	9	11
1	58.9	67.8	64.4	66.7	71.1	**75.6**
2	68.9	66.7	63.3	68.9	73.3	**74.4**
3	75.6	75.6	74.4	76.7	**80.0**	76.7
4	71.1	76.7	76.7	**80.0**	77.8	**80.0**
5	65.6	71.1	73.3	**77.8**	**77.8**	73.3
10	70.0	77.8	80.0	81.1	**83.3**	77.8

つ設定し，かつ，$k = 1$ とした場合，上記の性別推定法は文献 [1] の著者推定法となる．

3.2 実験

3.2.1 実験 1

まず，推定対象テキストの著者が参照著者集合に含まれるという条件下で実験を行う．具体的には，以下のような設定で推定精度を 3 分割交差検定により求める．

推定対象テキスト
: エッセイコーパスにおける各著者のテキストデータ 3 冊から 1 冊を選び，先頭から n パッセージ $(1 \leq n \leq 10)$ を推定対象テキストとして用いる．

参照テキスト集合
: 上記で除いた各著者のエッセイ集 2 冊 (20 パッセージ) を用いる．総参照テキスト数は著者数と同数のため 30 となる．

推定精度は，推定対象テキストの大きさ n，および，多数決を行なう参照テキスト数 k に依存する．実験結果を表 1 に示す．

このような実験設定では，推定対象テキスト Q に対して，それと同一著者の参照テキスト T_i が最も類似するテキストになれば，正しい性別が得られる ($k = 1$ の場合)．つまり，著者推定によって，性別推定を実現できる．事実，著者推定は高い精度で実現できる [1] ため，$k = 1$ における精度は非常に高く，推定対象テキストが 2 パッセージのとき，性別推定精度は 90% を超える．しかし，k を大きくするにつれ，別の著者のテキストも多数決の対象となるため，精度は下がることになる．

3.2.2 実験 2

性別推定が必要とされる状況下では，実験 1 のように推定対象テキストの著者が参照著者集合に含まれることは想定しにくい．そこで，実験 2 では，推定対象テキストの著者が参照著者集合に含まれないという設定で，実験を行う．具体的には，以下のような設定で推定精度を求める．

推定対象テキスト
: エッセイコーパスから著者 1 人を選び，エッセイ集 3 冊それぞれに対し，先頭から n パッセージ $(1 \leq n \leq 10)$ を推定対象テキストとして用いる．

参照テキスト集合
: 上記の著者を除いた 29 人のエッセイ集 3 冊のテキストデータを，それぞれ 1 冊単位のテキストデータに分割し用いる．すなわち，各参照テキストのサイズは 10 パッセージ，総参照テキスト数は 29 人 $\times$ 3 冊 $= 87$ となる．

実験結果を表 2 に示す．実験 1 の結果と比較し，実験 2 では精度が大きく下がることが分かる．つまり，推定対象テキスト Q の著者が参照著者集合に含まれる否かという条件は，性別推定の精度を大きく左右する．それゆえ，性別推定の実験においては，どちらの条件で実験しているかを必ず明示する必要がある．

3.2.3 実験 3

実験 2 では，ひとつの参照テキストを，エッセイ集 1 冊 (10 パッセージ) で構成した．これに対して，参照テキストを著者 1 人単位 (30 パッセージ) で構成した場合，推定精度は実験 2 の推定精度より低い値となった (詳細は省略する)．この事実は，参照テキストの数を増やすことにより，精度が向上する可能性があることを示唆する．

表 3: 実験 3 の結果

	k							
n	1	3	5	7	9	11	13	15
1	64.4	**75.6**	73.3	70.0	**75.6**	74.4	**75.6**	70.0
2	64.4	67.8	68.9	67.8	72.2	73.3	75.6	**76.7**
3	75.6	76.7	74.4	74.4	77.8	**81.1**	80.0	83.3
4	66.7	75.6	80.0	78.9	**81.1**	80.0	80.0	78.9
5	66.7	77.8	73.3	76.7	**81.1**	**81.1**	77.8	72.2
10	75.6	83.3	**85.6**	**85.6**	82.2	**85.6**	82.2	76.7

そこで，実験 3 では，実験 2 の設定に，参照テキストを追加し，精度がどう変化するかを調べた．具体的には，次のような実験設定を用いた．

推定対象テキスト
: 実験 2 と同様のテキストを用いる．

参照テキスト集合
: 実験 2 で用いた参照テキスト集合に，BCCWJ から，NDC が 914（エッセイ），かつ，テキストサイズが 5,000 字以上のもの[1] を，男女それぞれ 20 テキスト加える．総参照テキスト数は，87 + 40 = 127 となる．

実験結果を表 3 に示す．この結果から，参照テキスト数を増やすことにより精度が向上することが確かめられた．実験 2 では，平均すると $k = 9$ のとき最も精度が良いが，本実験では $k = 11$ のとき最も精度が良い．すなわち，参照テキスト数が異なれば，精度の良い k も異なる．

一方，推定対象テキストのパッセージ数 n が異なれば，精度の良い k も異なる．この表において，n が小さい場合は k は大きいほうが精度が良く，逆に，n が大きい場合は k は小さいほうが精度が良い，という傾向が見られる．これは，n が小さい場合は類似テキストの推定精度の信頼が低下するため，より多くの参照テキストを多数決の対象に含めた方が高い精度が得られるが，n が大きい場合は少数の信頼できる類似テキストのみを多数決の対象としたほうが高い精度が得られるからだと考えられる．

4 二値分類器を用いた性別推定

3 節では k 近傍法を用いた性別推定を実現した．これに対し本節では，二値分類器を用いて著者の性別を推定する方法について検討する．

[1] テキストサイズが 5,000 字以上のものに限定したのは，実験 2 で用いた参照テキストが約 10,000 字であり，大きくサイズの異なるものは不適切と考えたためである．

4.1 方法

性別は男女の 2 つの値をとるので，二値分類器を構成して性別を推定するという方法は，素直なアプローチである．事実，これまでの性別推定の研究では，二値分類器を用いる方法が主流である．具体的には，次のような手順となる．

1. 2 つのテキスト集合 $\mathbf{T_M}$, $\mathbf{T_F}$ を準備する．ここで，$\mathbf{T_M}$ は男性著者のテキスト集合，$\mathbf{T_F}$ は女性著者のテキスト集合である．これらが参照テキスト集合に相当する．
2. $\mathbf{T_M}$ と $\mathbf{T_F}$ から (なんらかの方法で) 二値分類器を構成する．
3. 推定対象テキスト Q が与えられる．
4. Q を二値分類器に入力し，性別を得る．

ここでは，どのような方法で二値分類器を構成するかが問題となる．本研究では，SVM を用いて二値分類器を構成する．SVM の学習に用いる素性としては，以下の 2 つのいずれかを用いる．

有効文字 **bigram**
: ひらがな，カタカナ，JIS 第一水準の漢字からなる文字 bigram を有効文字 bigram とし，素性はこの生起確率とする．これは，3 節で述べた手法で用いる文字 bigram と同じである．総素性数は 9,809,424 となる．

品詞 **bigram**
: 形態素解析で得られた品詞，活用型，活用形を 1 セットとし，素性はこの bigram とする[2]．解析器には，MeCab + IPAdic を用いる．記号は除いたため，総素性数は 184,041 となる．

SVM の実装として LIBLINEAR を用いる．カーネルは線形カーネル，コストマージンパラメータは最良の結果となるものを用いる．

[2] 先の研究 [2][3] では，品詞 n-gram が性別推定に有用であると示されている．

表 4: 実験 1b，および実験 2b の結果

	素性	n			
		1	3	5	10
1b	有効文字 bigram	83.7	–	**90.6**	90.0
	品詞 bigram	69.8	–	79.4	**85.6**
2b	有効文字 bigram	69.4	72.0	**73.3**	65.6
	品詞 bigram	56.6	59.7	**60.6**	56.7

4.2 実験

4.2.1 実験 1b

3.2.1 節の実験 1 に対応する実験を二値分類による推定法を用いて行う．エッセイコーパスにおける著者 1 人のテキストデータのうち，2 冊 (20 パッセージ) を学習，1 冊 (10 パッセージ) をテストに用いる．n パッセージ（$n = 1, 5, 10$）を 1 インスタンスとして，学習，テストに用いる．学習インスタンス数はそれぞれ，600，120，60 となる．それぞれの素性を用いた結果を表 4 に示す．

この結果より，有効文字 bigram を素性に用いた推定精度は，品詞 bigram を用いた場合に比べ高く，n が小さいほど精度の差は顕著であることが分かる．実験 1 の結果と比べると，両素性とも精度は低い．有効文字 bigram を素性として用いた場合，$n = 1$ では実験 1 に比べ大きな差はないが，$n = 5, 10$ と大きくすると，精度の差は大きくなる．$n = 10$ においてもそれほど精度が上がらないのは，学習インスタンス数が少なくなるためと考えられる．

4.2.2 実験 2b

3.2.2 節の実験 2 に対応する実験を二値分類による推定法を用いて行う．エッセイコーパスにおける著者 1 人のテキストデータをテストに用い，残り 29 人のテキストを学習に用いる．n パッセージ（$n = 1, 3, 5, 10$）を 1 インスタンスとして，学習，テストに用いる．学習インスタンス数はそれぞれ，870，290，174，87 となる．それぞれの素性を用いた結果を表 4 に示す．

実験 1b の結果と同様，有効文字 bigram を素性に用いた推定精度は，品詞 bigram を用いた場合に比べ高い．実験 2 と比べると，両素性とも精度は低い．

以上より，今回の実験では，二値分類器を用いた推定法は，k 近傍法を用いた推定法に比べ，推定精度が低いという結果となった．

5 関連研究

日本語を対象とした性別推定の研究に，池田ら [4] の blog を対象とした研究がある．素性に機能語，一人称，形態素を用いた二値分類器により性別を推定しており，最大で 88.9%の推定精度を得ている．

日本語以外では，Koppel ら [2] が，機能語，品詞 n-gram を用いた重み付けにより性別推定を実現している．BNC を用いた実験では，ノンフィクションを対象としたときの精度は 82.6%である．E メールを用いた Corney ら [3] による実験では，SVM を用い，素性に単語や文の長さ，機能語，HTML タグなどを使用した場合，7 割程度の推定精度を得ている．

いずれの研究においても，実験において，推定対象テキスト Q の著者が参照著者集合に含まれているかどうかの記述はない．

6 おわりに

本論文では，エッセイコーパスを用いた性別推定について述べた．推定対象テキストの著者が参照著者集合に含まれる場合と含まれない場合それぞれで実験を行ったとき，推定精度に大きな差があるという結果が得られた．推定法には，k 近傍法を用いる方法と，二値分類器を用いる方法を実装し，前者のほうが精度が良いという結果が得られた．

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] 石田 将吾, 佐藤 理史: 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 (2010)

[2] Moshe Koppel, Shlomo Argamon, Anat Rachel Shimoni: Automatically Categorizing Written Texts by Author Gender, In 18th Annual Computer Security Applications Conference (2002)

[3] Malcolm Corney, Olivier de Vel, Alison Anderson, George Mohay: Gender-Preferential Text Mining of E-mail Discourse, Literary and Linguistic Computing, 17(4), pp.401-412 (2002)

[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 (2006)