Vol.2010-NL-198 No.6
2010/9/17

エッセイコーパスを用いた日本語テキストの著者推定

石 田 将 吾[†1]　佐 藤 理 史[†1]

本論文では，新たに編纂したエッセイコーパスを用いた著者推定実験の結果について報告する．このエッセイコーパスは，30 人の職業作家の 90 冊のエッセイ集から抽出したテキストから構成されており，1 人当たり約 3 万字（約 1,000 字 ×10ヶ所 ×3 冊）を収録している．文字 bigram 言語モデルを利用した著者推定法の精度は，5,000 字の推定用テキストを用いた場合，97.8%であった．また，推定用テキストとして，1ヶ所から抽出した 1,000 字を用いた場合は 74.4%，5ヶ所から抽出した 200 字を併合した 1,000 字を使った場合は 84.9%と，推定精度が 10 ポイント以上異なることを明らかにした．

Authorship Attribution with Essay Corpus

Shogo Ishida[†1] and Satoshi Sato[†1]

This paper reports experimental results of authorship attribution with a newly compiled Essay Corpus. The Essay Corpus consists of 900,000 characters, extracted from 90 essay books of 30 professional writers; from each book, ten passages of 1,000 characters are extracted. We have obtained 97.8% accuracy of authorship attribution when we use 5,000 characters as a test text to be identified. In addition, we have observed that the accuracy varies according to the number of passages from which the test text is made, even if the text size is fixed.

1. はじめに

著者推定とは，著者が不明のテキストが与えられたとき，その著者が誰であるかを推定する問題である．その研究は長い歴史を持ち，計算機の登場以前から様々な手法が提案されてきた[1)]．その応用領域は，計算機やウェブの発達に伴って拡大しており，最近ではブログの著者推定，同一著者判定などが研究されている[2)]．

一般に，著者推定は，次のような問題設定をとる．まず，著者集合 A を定める．次に，その集合に含まれるそれぞれの著者 a_i に対して，その著者が書いたテキスト T_i を準備する．本論文では，これを**規準テキスト**と呼ぶ．そして，あるテキスト Q（これを**推定用テキスト**と呼ぶ）が与えられたとき，そのテキストの著者を，著者集合 A の中から選ぶ．このような問題設定は，以下のように整理される．

与えられるもの

（1） 著者集合: $A = \{a_1, a_2, \ldots, a_n\}$

（2） 規準テキスト集合: $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$

（3） 推定用テキスト: Q

推定すべきもの　推定用テキスト Q の著者 $a_i \in A$

著者推定の研究は，英語テキストを対象としたものが多い．最近では，数千人の著者を対象にした実験[3)] や，少量の推定用テキスト及び規準テキストを用いた実験[4)] が行われている．

これに対し，日本語テキストを対象とした著者推定の研究はそれほど活発には行なわれていない．松浦ら[5)] は，8 人の作家によって書かれた 10,000〜30,000 字の規準テキストと，それと同規模の推定用テキストを用いた実験について報告している．また，西村ら[6)] は，Yahoo!知恵袋に投稿された 10 人のテキストを用いた実験を行っている．

我々は，今回，職業作家 30 人のエッセイを収録したエッセイコーパスを編纂した．エッセイを収録対象とした理由は，小説や専門書などに比べ，書き手の個性が現れやすいと考えたからである．

本論文では，このエッセイコーパスを用いた著者推定の実験結果について報告する．具体的には，著者数や推定用テキストサイズを変化させたときの推定精度の変化，および，推定用テキストの作成法と推定精度の関係について報告する．

2. エッセイコーパス

テキストには書き手の個性が現れる．そのような個性を研究すべく，我々はエッセイを収録対象としたコーパスを設計し，次のような手順で編纂した．

（1） 著者 10 人 (男性 5 人，女性 5 人) を選定する．

（2） それぞれの著者に対して，出来るだけ題材に偏りのないエッセイ集（単行本）を 3 冊

†1 名古屋大学大学院工学研究科電子情報システム専攻
Department of Electrical Engineering and Computer Science, Guraduate School of Engineering, Nagoya University

Vol.2010-NL-198 No.6
2010/9/17

表 1 エッセイコーパスに収録した 30 人の著者リスト
Table 1 Thirty authors in Essay Corpus

G_1	G_2	G_3
司馬遼太郎（1923）	遠藤周作（1923）	北杜夫（1927）
五木寛之（1932）	開高健（1930）	渡辺淳一（1933）
椎名誠（1944）	林望（1949）	阿刀田高（1935）
村上龍（1952）	辻仁成（1959）	大江健三郎（1935）
原田宗典（1959）	大槻ケンヂ（1966）	中島らも（1952）
岡部伊都子（1923）	佐藤愛子（1923）	森瑤子（1940）
森村桂（1940）	阿川佐和子（1953）	内館牧子（1948）
林真理子（1954）	群ようこ（1954）	柴門ふみ（1957）
中村うさぎ（1958）	江國香織（1964）	酒井順子（1966）
さくらももこ（1965）	よしもとばなな（1964）	鷺沢萌（1968）

選ぶ.

(3) 選んだエッセイ集 1 冊につき，等間隔に 10 箇所から約 1,000 字ずつ抽出し，電子化する.

この手順を 3 回繰り返し，総計 30 人の著者からなるコーパスを編纂した．収録著者名を **表 1** に示す．この表の括弧内の数字は各著者の生年を表している．G_1，G_2，G_3 のグループは，上記の手順のサイクルに対応しており，後述する実験で用いる．

以下では，次の用語を用いる．抽出した約 1,000 字単位のテキストを **パッセージ**と呼ぶ．1 冊の本から抽出した前半 5 パッセージ，および，後半 5 パッセージを，それぞれ**ユニット**と呼ぶ．著者 1 人当たりのユニット数は，2 ユニット × 3 冊 = 6 ユニット，文字数は，約 1,000 字 ×5 パッセージ ×2 ユニット ×3 冊 = 約 30,000 字である．

3. 文字 bigram 言語モデルを用いた著者推定

3.1 文字 bigram 言語モデルの構築

ここでは，著者集合 A に含まれる各著者 a_i に対して，文字 bigram 言語モデル M_i を構築する.

テキスト中に現れる文字のうち，ひらがな，カタカナ，JIS 第一水準の漢字のみを有効文字とし，それ以外の数字や記号，アルファベット等は全て無視する．このとき，有効文字の総数（異なり）は 3,132 となる．著者 $a_i(a_i \in A)$ の言語モデル M_i における有効文字 bigram x_jx_k の生起確率 $P_i(x_k|x_j)$ を，次式で求める.

$$P_i(x_k|x_j) = \frac{f(x_jx_k, T_i)}{f(x_j*, T_i)} \tag{1}$$

ここで，x は有効文字，$*$ は任意の有効文字を表す．すなわち，文字 bigram x_jx_k は，x_j と x_k の両方が有効文字であるもののみを用いる．これを**有効文字 bigram** を呼ぶ．有効文字 bigram の総数（異なり）は $3,132^2 = 9,809,424$ となる．T_i は著者が a_i である規準テキストを表し，$f(x_jx_k, T_i)$ は規準テキスト T_i における有効文字 bigram x_jx_k の出現回数を表す.

式 (1) は，右辺の分子が 0 の場合，生起確率が 0 となる．分母が 0 の場合は生起確率が計算できない．このため，次の 2 つの補正法のいずれかを採用する.

補正法 1：Good-Turing 推定法を用いた補正

Good-Turing 推定法[7] は，次式を用いて出現回数 f を f_{GT} に補正する方法である.

$$f_{GT} = (f+1)\frac{N_{f+1}}{N_f} \tag{2}$$

ここで，N_f は，テキストに f 回出現する有効文字 bigram の異なり数を表す．N_0 は，テキスト中に出現しない有効文字 bigram の異なり数を表す.

f が大きいと N_f が 0 となる場合があるため，実際の補正には，次式を用いる.

$$f_{GT} = \begin{cases} (f+1)\frac{N_{f+1}}{N_f} & \text{if } 0 \le f \le 3 \\ f-1 & \text{if } f \ge 4 \end{cases} \tag{3}$$

上式で計算される f_{GT} を用いて，言語モデル M_i における有効文字 bigram x_jx_k の生起確率を次式で計算する.

$$\hat{P}_i(x_k|x_j) = \frac{f_{GT}(x_jx_k, A_i)}{f_{GT}(x_j*, A_i)} \tag{4}$$

補正法 2：小さな定数を用いた補正

この補正は，コーパスに出現しなかった有効文字 bigram の出現回数を，小さな正の定数で置き換えるものである．具体的には，

(1) 有効文字 bigram x_j* が出現しなかった場合，x_j* は 10 回出現し，x_jx_k は 0.1 回出現したものとみなす.

(2) 有効文字 bigram x_j* は出現したが，x_jx_k は出現しなかった場合，x_jx_k が 0.1 回出現したものとみなす.

すなわち，言語モデル M_i における有効文字 bigram x_jx_k の生起確率を次式で計算する.

Vol.2010-NL-198 No.6
2010/9/17

$$\hat{P}_i(x_k|x_j) = \begin{cases} \dfrac{0.1}{10} = 0.01 & \text{if } f(x_j*, T_i) = 0 \ \wedge \ f(x_j x_k, T_i) = 0 \\ \dfrac{0.1}{f(x_j*, A_i)} & \text{if } f(x_j*, T_i) = 0 \ \wedge \ f(x_j x_k, T_i) > 0 \\ \dfrac{f(x_j x_k, T_i)}{f(x_j*, T_i)} & \text{otherwise} \end{cases} \quad (5)$$

3.2 尤度計算

推定用テキスト Q の著者を求めるために，各言語モデル M_i に対する尤度 $L(M_i|Q)$ を，次式を用いて計算する．

$$L(M_i|Q) = \sum_{x_j x_k \in Q} f(x_j x_k, Q) \log \hat{P}_i(x_k|x_j) \quad (6)$$

こうして得られる尤度のうち，最大の尤度をとる言語モデル M_i を求め，これに対応する著者 a_i を推定結果として出力する．

4. 実　　験

本節では，著者数，推定用テキストのサイズ，推定用テキストの作成法の 3 つが，著者推定精度にどのように影響するかを調べる．

4.1 実　験　1

実験 1 では，著者数及び推定用テキストのサイズが著者推定精度にどのように影響するかを調べる．

4.1.1 方　　法

2 節で述べたように，エッセイコーパスでは，著者 1 人に対するテキストデータが 6 ユニットから構成されている．このうち 5 ユニットを規準テキストとして利用し，1 ユニットを推定用ユニットとして利用する．

著者数 10 人，20 人，30 人のそれぞれの場合に対して，推定精度を測定する．著者数 10 人の場合は，表 1 の G_1，G_2，G_3 のそれぞれのグループに対し，推定精度を 6 分割交差検定により求め，得られた精度の平均値を計算する．著者数 20 人の場合は，G_1+G_2，G_2+G_3，G_3+G_1 の 3 種類の著者集合に対して推定精度を 6 分割交差検定により求め，得られた精度の平均値を計算する．著者数 30 人の場合は，$G_1+G_2+G_3$ に対して著者推定精度を 6 分割交差検定により求める．

推定用ユニットは，5 パッセージから構成される．このうち，先頭から k パッセージ（$1 \leq k \leq 5$）を推定用テキストとして用いる．すなわち，長さが異なる 5 種類の推定用テキストに対して，著者推定を行う．

表 2　著者推定結果（実験 1）
Table 2　Result of Experiment 1

	各著者数における精度（%）					
	Good-Turing 推定法			小さな定数		
推定用テキストサイズ	10 人	20 人	30 人	10 人	20 人	30 人
1 パッセージ	78.9	70.3	65.0	82.2	77.8	74.4
2 パッセージ	88.9	81.4	77.2	93.3	89.7	87.2
3 パッセージ	92.8	88.6	86.1	97.8	95.6	94.4
4 パッセージ	94.4	90.3	87.2	98.3	96.4	95.6
5 パッセージ	95.5	93.0	90.6	98.9	98.1	97.8

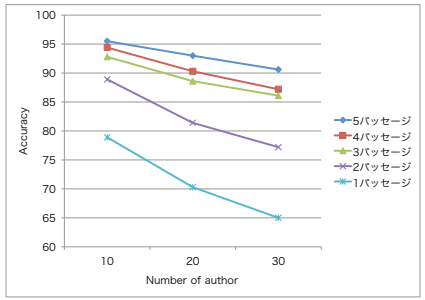

図 1　Good-Turing 推定法による補正を用いた結果
Fig. 1　Result of authorship attribution using Good-Turing smoothing

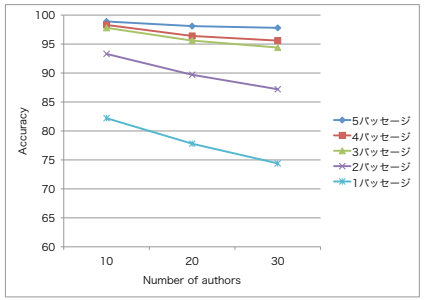

図 2　小さな定数による補正を用いた結果
Fig. 2　Result of authorship attribution using smoothing with small constant values

4.1.2 結　　果

実験結果を**表 2** に示す．著者推定精度のチャンスレベルは，著者数 10 人のとき 10%，20 人のとき 5.0%，30 人のとき 3.3%である．

まず，3.1 節で述べた 2 種類の補正法を比較する．2 種類の補正法の推定精度をグラフ化したものを**図 1** および**図 2** に示す．この 2 つのグラフを見比べることにより，小さな定数による補正（図 2）の方が，すべての場合において精度が高く，かつ，著者数の増加に伴う精度の低下も小さいことがわかる．この結果に基づき，補正法としては，小さな定数による補正を採用することとし，以下では，この補正法を用いた場合の実験結果について議論する．

次に，著者数の増加に伴う精度の変化に注目する．著者数を 10 人から 20 人に増やしたときの精度の低下に比べ，20 人から 30 人に増やしたときの精度の低下は小さい．また，テ

Vol.2010-NL-198 No.6
2010/9/17

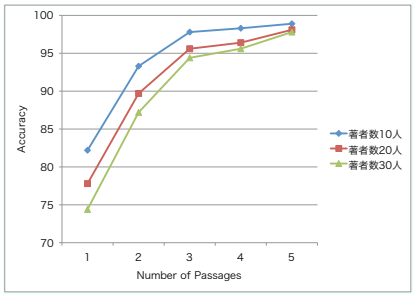

図 3 パッセージ数と著者推定精度の関係
Fig. 3 The number of passages vs. accuracy

キストサイズが大きいほど，著者数を増やしたときの精度の低下は小さい．推定用テキストサイズが 1 パッセージの場合，10 人から 20 人に増やしたときの精度の低下は 4.4 ポイント，20 人から 30 人に増やしたとき精度の低下は 3.4 ポイントである．これに対して，推定用テキストサイズが 5 パッセージの場合，精度の低下はそれぞれ 0.8 ポイント，0.3 ポイントである．

推定用テキストのパッセージ数と推定精度の関係をグラフ化したものを**図 3** に示す．パッセージ数を 1 から 2 に増やしたときの精度の上昇は，著者数 10 人の場合は 11.1 ポイント，著者数 30 人の場合は 12.8 ポイントと，かなりの大きな精度上昇が見られる．しかしながら，その精度上昇は，パッセージ数を増やすにつれて減少する．

ここで一つの疑問が生じる．上記の精度の上昇は，テキストサイズ（文字数）の増加によるものなのか，それともパッセージ数の増加によるものなのか，という疑問である．エッセイコーパスでは，オリジナルテキストから抽出する連続したテキスト（パッセージ）は約 1000 字で固定されており，それ以上の長さのテキストを推定用テキストとする場合は，複数のパッセージから構成されるテキストを使用せざるを得ない．それ故，この実験だけからは，上記の疑問に答えることができない．

4.2 実　験　2

上記の疑問に答えるために，以下のような実験を行った．

4.2.1 方　　法

実験 1 と同様の 6 分割交差検定により著者 30 人の場合の著者推定精度を求める．但し，1 つの推定用ユニットから，次のような方法で約 1,000 字の推定用テキストを複数作成する．

表 3 実験 2 の組み合わせと試行回数
Table 3 The combination and the number of trials in Experiment 2

種類数 n	組み合わせ	試行回数 K_n	生成する推定用テキスト数
1	${}_5C_1$	1	5
2	${}_5C_2$	1	10
3	${}_5C_3$	1	10
4	${}_5C_4$	2	10
5	${}_5C_5$	10	10

(1) 推定用テキストの作成に使用するパッセージの種類数 n（$1 \leq n \leq 5$）を定める．
(2) 5 つのパッセージから n 個のパッセージを選ぶ．この操作を，全ての可能な組み合わせに対して行う．
(3) 選んだ n 個のパッセージをそれぞれ n 分割し，各パッセージからランダムに 1 つずつ選び，これらを繋ぎ合わせる．これを推定用テキストとする．この操作を K_n 回行う．

この手続きにより，1 つの推定用ユニットから，ある n に対して ${}_5C_n \times K_n$ 個の推定用テキストが生成される．各 n に対する K_n の値，および，生成する推定用テキストの数を**表 3** に示す．K_n の値は，$n = 1$ の場合を除いて生成する推定用テキスト数が同じ数となるように定めた．なお，$n = 1$ の場合は，5 つのパッセージをそれぞれ 1 つずつ推定に用いる場合と同じである．最終的に，使用するパッセージ数 n に対して平均精度を計算する．

4.2.2 結　　果

実験結果を**表 4**，および，**図 4** に示す．表 4 の括弧内の数字は，使用するパッセージの種類数を 1 増やしたときの精度の増加を示している．

この表より，種類数の増加に従い，著者推定精度は上昇することが分かる．精度の上昇は，種類数を 1 から 2 に増やしたときが一番大きく（5.1 ポイント），それ以降の上昇は，種類数が増えるにつれて緩やかになっている．

この実験結果は，次のことを示している．

(1) 推定用テキストのサイズが同じでも，その推定用テキストをいくつのパッセージから作成したかが異なれば，著者推定精度は異なる．言い替えるならば，推定用テキストのサイズと推定精度の関係を求めるためには，テキストの作成法を固定する必要がある．
(2) 複数のパッセージから推定用テキストを作成する方が，1 つのパッセージから作成するよりも高い推定精度が得られる．

表 4 著者推定結果（実験 2）
Table 4 The result of Experiment 2

パッセージの種類数	精度（%）	
1	74.6	
2	79.7	(+5.1)
3	82.9	(+3.2)
4	83.0	(+0.1)
5	84.9	(+1.9)

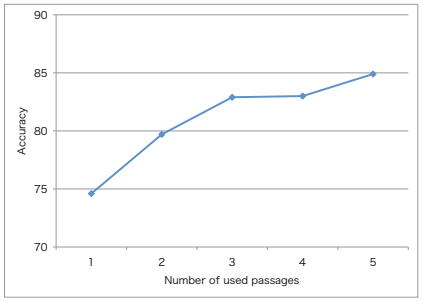

図 4 著者推定結果（実験 2）
Fig. 4 The result of Experiment 2

表 5 著者推定結果（実験 3）
Table 5 The result of Experiment 3

テキストサイズ	精度（%）
1,000 字	84.9
1,250 字	88.0
1,665 字	91.0
2,500 字	94.7
5,000 字	97.8

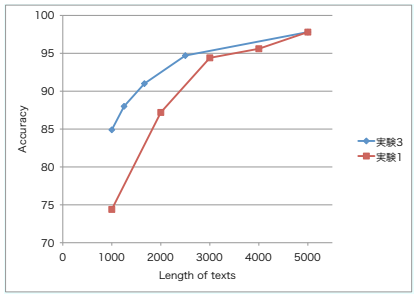

図 5 著者推定結果（実験 3）
Fig. 5 The result of Experiment 3

連続したテキストで構成されるパッセージは，ある特定のトピックに対する記述とみなすことができる．推定用テキストを 1 つのパッセージで構成する場合，トピックに対する特徴が著者の特徴を抑え，支配的になる可能性がある．一方，複数のパッセージから推定用テキストを作成する場合，それぞれのパッセージのトピックに対する特徴は相対的に弱まり，著者の特徴が際立ってくると考えられる．複数パッセージから推定用テキストを作成した場合の精度向上は，このような理由によるものと考えられる．

4.3 実　験　3

実験 2 から，推定用テキストの作成法が著者推定精度に影響することがわかった．すなわち，実験 1 の結果は，この影響を受けていることになる．そこで，実験 3 では，推定用テキストの作成法を固定し，推定用テキストのサイズと推定精度の関係を調べる．

4.3.1 方　　法

実験 1 と同様に，6 分割交差検定により著者 30 人の場合の著者推定精度を求める．但し，推定用テキストの作成に使用するパッセージの種類数を 5 に固定したまま，推定用テキストのサイズのみを変更する．

具体的には，推定用ユニットから，次のような手順で推定用テキストを作成する．

(1) パッセージの分割数 m（$1 \leq m \leq 5$）を定める．
(2) 推定用ユニットに含まれる 5 パッセージをそれぞれ m 個に分割する．
(3) 各パッセージから $1/m$ のパッセージをランダムに 1 つずつ選び，繋ぎ合わせたものを推定用テキストとする．これを 10 回繰り返す．

なお，$m = 1$ の場合は，5 パッセージ全てを推定用テキストとして用いることと同じである．一方，$m = 5$ の場合は，実験 2 で行った，5 種類のパッセージのそれぞれ 1/5 を繋ぎ合わせた場合と同じである．一つの推定用テキストのサイズは，約 1,000 字 $\div m \times 5$ パッセージ $= 5000/m$ 字となる．

4.3.2 結　　果

実験結果を**表 5**，および，**図 5** に示す．なお，図 5 では，実験 1 における著者 30 人での結果も，“実験 1”として示した．

この図より，テキストサイズを増加させることにより，著者推定精度が向上することがわかる．同時に，精度向上は，テキストサイズの増加に伴って次第に減少傾向にあることがわかる．本実験において，1,000 字から 5,000 字に増やしたときの精度の上昇は 12.9 ポイントであった．

実験 1 では，1,000 字から 5,000 字に増やしたときの精度の上昇は 23.4 ポイントであった．この 2 つの実験の差，すなわち，10.5 ポイントが，推定用テキストの作成法の違いに起因する上昇分となる．言い替えるならば，実験 1 の精度上昇は，テキストのサイズの増加と推定用テキストの作成法の違い（使用するパッセージの種類数の増加）の 2 つの影響を受けていたということである．

5. 関 連 研 究

著者推定の研究は，計算機やウェブの発達による応用領域の拡大に伴って，様々な研究が行われている（**表 6**）．Koppel ら[3] は，数千人のブログを用いることで，大規模な著者集合・規準テキスト集合を構成し，これを用いた著者推定実験を行っている．Hirst ら[4] は少

Vol.2010-NL-198 No.6
2010/9/17

表 6 関連研究との比較
Table 6 The summary related work

論文	使用するテキスト	推定用テキストサイズ	規準テキストサイズ	著者数	精度（%）
Koppel[3]	ブログ	500words 以上	200 記事以上	数千	35
Hirst[4]	小説	200words	200words	2	92
Luyckx[9]	学生のエッセイ	280words	1120words	5	88
				10	82
				20	80
				50	60
				100	50
Peng[10]	新聞記事	900words	900words	10	90
西村[6]	Yahoo!知恵袋	1 記事	約 2,500 記事	10	94
提案手法	エッセイ	5,000 語	25,000 語	30	98

量の推定用テキスト及び規準テキストを用いた実験を行っており，Stamatatos[8] はテキストサイズに偏りがある場合の影響を調べている．著者数と著者推定精度の関係を示した研究としては，Luyckx ら[9] が，学生 145 人のエッセイから構成されるコーパスを用いて，著者数の変化に伴う精度の変化を示している．日本語テキストに対しては，松浦ら[5] が青空文庫から作成したコーパスを用いて，著者推定実験を行っており，西村ら[6] は，Yahoo!知恵袋のテキストを用いた実験を行っている．しかし，これらの研究は，推定用テキストの作成法が著者推定精度に与える影響について調査していない．複数のパッセージから推定用テキストを構成することにより，推定精度が向上することは，本研究によって初めて示された知見である．

6. おわりに

本論文では，新たに編纂したエッセイコーパスを用いた，著者推定実験の結果について述べた．文字 bigram 言語モデルを利用した著者推定法では，30 人の著者の著者集合を対象として 5,000 字の推定用テキストを用いた場合，97.8%の推定精度が得られた．また，推定用テキストの作成法が異なれば，推定精度が変化することを発見した．推定用テキストのサイズが同じ 1,000 字であっても，5ヶ所から抽出した 200 字を併合した 1,000 字の場合の推定精度は，1ヶ所から抽出した 1,000 字を用いた場合に比べ，10 ポイント以上高いという結果が得られた．

参 考 文 献

1) 村上征勝: 真贋の科学–計量文献学入門, 東京, 朝倉書店 (1994)
2) Efstathios Stamatatos: A survey of modern authorship attribution methods, Journal of the American Society for information Science and Technology, 60(3), pp. 538–556 (2009)
3) Moshe Koppel, Jonathan Schler, Shlomo Argamon, Eran Messeri: Authorship attribution with thousands of candidate authors, Proceedings of the 29th ACM SIGIR, pp.659–660 (2006)
4) Graeme Hirst and Ol'ga Feiguina: Bigrams of syntactic labels for authorship discrimination of short texts, Literary and Linguistic Computing, 22(4), pp.405–417 (2007)
5) 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, 情報処理学会自然言語処理研究会報告, NL Vol.137, No.1, pp.1–8 (2000)
6) 西村 涼, 渡辺 靖彦, 村田 真樹, 岡田 至弘: Yahoo!知恵袋に投稿されたテキストに対する著者判別, 言語処理学会第 15 回年次大会, pp.2-22 (2009)
7) Christopher D. Manning, Hinrich Schütze: Foundations of statistical natural language processing, Cambridge, Massachusetts, MIT Press (1999)
8) Efstathios Stamatatos: Author identification: Using text sampling to handle the class imbalance problem, Information Processing and Management, 44(2), pp.790–799 (2008)
9) Kim Luyckx, Walter Daelemans: Authorship attribution and verification with many authors and limited data, Proceedings of the 22nd International Conference on Computational Linguistics, pp.513–520 (2008)
10) Fuchun Peng, Dale Schuurmans, Shaojun Wang: Language and task independent text categorization with simple language models, Proceedings of HLT-NAACL, pp. 110–117 (2003)