語彙と文脈に着目した文学作品の著者推定

花畑圭佑　　青野雅樹
豊橋技術科学大学　情報・知能工学課程
hanahata@kde. cs. tut. ac. jp,　aono@cs. tut. ac. jp

1. はじめに

文章の著者推定問題[1]は，近年の SNS や Web ニュースサイトといったインターネットメディアの発達による電子テキストの急速な増加により，注目を集めているタスクである．匿名の文章の著者の推定，盗作や剽窃の検出，類似性が高い著者を発見することによる小説作品の推薦など，様々な分野に応用ができる．国際ワークショップである PAN@CLEF では，著者推定問題をはじめとする著者分析タスクが大きく取り上げられている．こういった背景から，著者推定問題は自然言語処理の分野で重要なテーマのひとつとなっている．

しかし，日本の文学作品を対象とした著者推定に関する研究は，我々が調査した範囲では多くは行われていない．そこで本研究では，文学作品の著者の語彙と文脈の特徴に着目した著者推定手法を提案する．

2. 関連研究

PAN@CLEF における著者推定を含む著者分析タスクは，近年活発に研究が行われている．一対の文章を与えられ，それらが同じ著者によって書かれたものか検証する著者照合タスクは Douglas[2], Alberto ら[3]等によって行われている．複数の文書を与えられ，著者ごとにグループ化する著者クラスタリングタスクは Houda[4], Yasmany ら[5]等によって行われている．また，現在論文は公開されていないが，PAN@CLEF 2018[1]では，一つの文書に複数のラベル付けがされている著者推定問題が公開されている．

小説作品に関する研究は，馬場ら[6]が作品のジャンルを推定する研究を行っている．松浦ら[7]は n-gram を用いた著者推定を行っている．

3. データセット

本研究では，青空文庫[8]にて公開されている文学作品を収集し，データセットとして使用する．青空文庫では新字新仮名，新字旧仮名，旧字旧仮名の 3 種類の仮名遣いの作品が公開されているが，本研究では新字新仮名の作品のみを使用した．新字新仮名の作品が 60 作品以上ある著者から 10 人を選択し，作品を収集した．選択した著者とその作品数は表 1 の通りである．

表 1 青空文庫データセット

著者名	作品数
芥川龍之介	163
太宰治	203
江戸川乱歩	79
泉鏡花	106
宮沢賢治	85
夏目漱石	83
坂口安吾	301
与謝野晶子	84
森鴎外	69
夢野久作	153

データの単位について，作品中で段落替えが発生するまでを 1 データとした．結果，合計 1, 315 件の作品から 277, 479 件の段落データを得た．以降，段落データのことを文章と呼ぶ．

4. 提案手法

文学作品には，著者のよく使用する語彙の特徴や，よく現れる文脈などの特徴が出ると考えられる．本研究では，語彙の特徴に着目した分類モデルと，文脈の特徴に着目した分類モデルの二つを訓練し，その出力をアンサンブルすることで，語彙と文脈両方の特徴に着目した分類モデルを提案する．提案手法の概要を図 1 に示す．

図 1　提案手法概要図

4.1 語彙特徴量による著者推定

著者ごとの語彙の特徴を学習するため，文章から Bag of Words(BoW)素性を得る．全文章には 105609 種類の語彙が存在した．学習コストの削減，重要度の低い素性の削除のため，次元削減を行う．全文章のうち，

20%以上の文書に現れる一般的な語彙と, 0. 005%以下の文章にのみ現れる希少な語彙を素性から削除し, 26093 次元の Bag of Words 素性を得た. 得た素性を, 多層パーセプトロンを用いて学習し, 著者の推定を行う. 図 2 に語彙特徴モデルのネットワーク構造を示す.

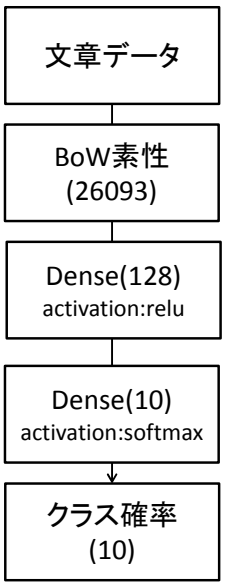

図 2　語彙特徴モデル

4. 2　文脈特徴量による著者推定

この節では, 文学作品の著者の, 文脈の特徴に着目した推定モデルを提案する. 文章を時系列データとして扱い, 再帰型ニューラルネットワークである LSTM を用いて著者推定を行う.

4. 2. 1 単語分散表現

訓練データの文学作品に登場する語彙の頻度上位 1 万単語を抽出し, インデックス化を行う. そのインデックスを用いて訓練データの文章に現れる単語を単語分散表現に変換する. 単語分散表現には, Facebook が公開[9]している, 事前学習された fastText[10]モデルを使用する. Common Crawl と日本語 Wikipedia をコーパスとして, CBOW を用いて, 次元数 300 で学習されている. 各データの長さを均一にするため, 単語列の最大長を 200 とする. こうして, 300×200 次元の時系列データを得る.

4. 2. 2 Long Short Term Memory (LSTM)

得られた時系列データを入力とし, LSTM を学習する. LSTM ユニットの出力は 128 次元とした. 出力層の活性化関数には softmax 関数を用いる. 図 4 に文脈特徴モデルのネットワーク構造を示す.

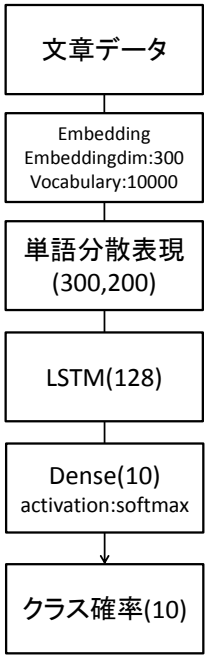

図 3　文脈特徴モデル

4. 3　アンサンブル学習

4. 1 節と 4. 2 節にて構築した二つのモデルからそれぞれクラス確率を得る. 得られたクラス確率の要素積をとり, 最も高いクラス確率の著者を, 文章の推定著者とする.

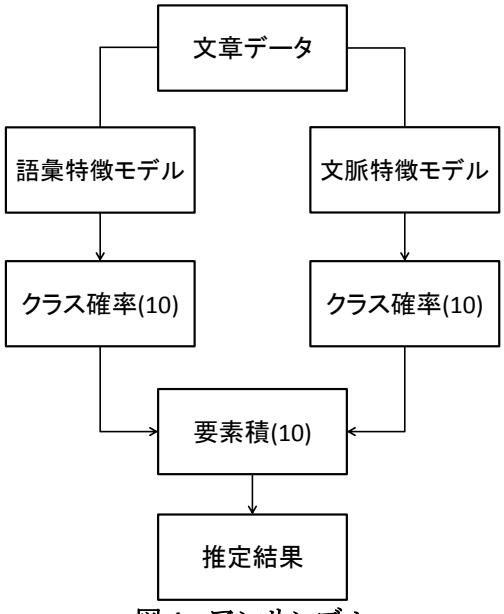

図 4　アンサンブル

5.　評価実験

ベースラインとして, 4. 1 節, 4. 2 節で構築した語彙特徴モデルと文脈特徴モデル単体での実験結果を用いる. 評価指標として accuracy を用いる.

5.1 データの分割

本実験では，収集した作品を 8:2 の比率でランダムに分割し，それぞれ訓練データと評価データとして使用した．実験に用いるデータセットの内訳は表 2 の通りである．

表 2　データセット

	訓練	評価	合計
作品数	1,052	263	1315
文章数	211,649	65,830	277,479

5.2 実験結果

ベースラインと提案手法を比較した結果を表 3 に示す．二つのモデルをアンサンブルする提案手法が，単独のモデルを使用する手法に比べ精度が向上することが確認できた．これは語彙の特徴と文脈の特徴の両方を効果的に分類に活用できたからだと考えられる． 各手法における混合行列を図 5 に示す．横軸が真値，縦軸が予測値を示す．対角線上に存在する数値が正しく推定できた数である．森鴎外の作品の正解率が低いことがわかるが，原因の解明には至っていない．

表 3　実験結果

手法	accuracy
ベースライン1(語彙モデル)	0.5620
ベースライン2(文脈モデル)	0.5962
提案手法	0.6440

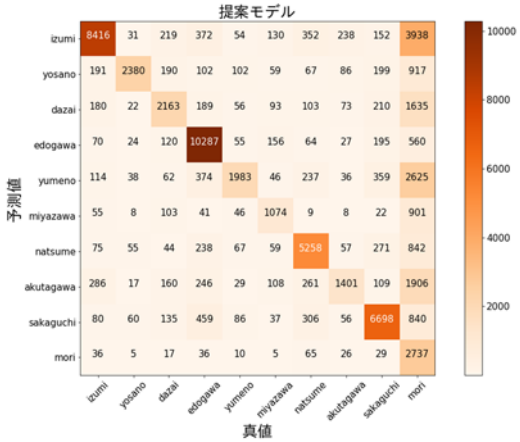

図 5　混合行列

5.3 考察

成功例と失敗例を以下に示す．

図 6 は評価データの一つであり，江戸川乱歩の「大金塊」という作品の一部である．ベースライン 2 の文脈モデルでは，この文章を宮沢賢治と誤分類したが，提案手法では江戸川乱歩と正しく分類した．この文章を観察すると，「はげしく」「たしかに」「おそろしい」というように，全体的に平仮名の語彙が多いことがわかる．そこで，全データセットにおける，語彙「おそろしい」の登場する回数を調査した．数え上げた結果は表 4 の通りである．表 4 から，多くの著者では「おそろしい」は数回しか登場しておらず，江戸川乱歩の作品中には 873 回と，非常に偏って登場している語彙であることがわかる．こういった語彙の特徴を提案手法ではとらえることができたのだと考えられる．

図 7 は評価データの一つであり，夢野久作の「押絵の奇蹟」という作品の一部である．ベースライン 2 の文脈モデルでは，芥川龍之介と誤分類したが，提案手法では夢野久作と正しく分類した．この文章においても夢野久作特有の語彙が存在すると考え，各語彙の搭乗頻度を調査したところ，語彙「御座い」の部分に大きく特徴が表れていることが確認できた．数え上げた結果は表 5 の通りである．先ほどの例と同様に，「御座い」は夢野久作の作品に非常に多く登場する語彙であり，その特徴を提案手法でとらえることができたと考えられる．

図 8 は評価データの一つであり，太宰治の「断崖の錯覚」という作品の一部である．ベースライン 2 の文脈モデルでは太宰治と正しく分類したが，提案手法では坂口安吾と誤分類した．表 6 の通り，語彙「テエブル」は芥川龍之介の作品に最も多く登場しており，そこに大きく影響を受けた結果誤分類してしまったのだと思われる．こういった問題の改善案として，クラス確率の要素積をとる際の重みを変更するといったことが挙げられる．

> 水の動き方はだんだんはげしくなってきました．たしかに一方にむかって流れているのです．**おそろしい**いきおいで流れているのです．ふたりはまた手を取りあって，流されまいと，ぎゃくに泳いでみましたが，だめでした．急流のような早い流れにさからうことはできません．

図 6　成功例 1

表 4　「おそろしい」の登場回数

著者名	登場回数
芥川龍之介	0
太宰治	100
江戸川乱歩	873
泉鏡花	1

宮沢賢治	1
夏目漱石	0
坂口安吾	2
与謝野晶子	2
森鴎外	3
夢野久作	14

その阿古屋の琴責めの五人組の人形が，柴忠さんの家の小さな本檜舞台に飾られました時の見物といったら，それは大変だったそうで**御座い**ます．申すまでもなくその時はお父様も，・・・お知り合いのお節句客の応対だけでも柴忠さんは眼がまわるほど，お忙がしかったそうで**御座い**ます．そうして・・・

図 7　成功例 2

表 5　「御座い」の登場回数

著者名	登場回数
芥川龍之介	1
太宰治	4
江戸川乱歩	13
泉鏡花	0
宮沢賢治	0
夏目漱石	39
坂口安吾	0
与謝野晶子	5
森鴎外	0
夢野久作	1641

まったく大人のような図太さで，私はグラスをカウンタア･ボックスの方へぐっと差しだした．日本髪の少女は，枯れかけた，鉢の木の枝をわけて，私の**テエブル**に近寄った．

図 8　失敗例

表 6　「テエブル」の登場回数

著者名	登場回数
芥川龍之介	91
太宰治	35
江戸川乱歩	0
泉鏡花	0
宮沢賢治	0
夏目漱石	0
坂口安吾	0
与謝野晶子	5
森鴎外	2
夢野久作	1

6.　おわりに

本研究では，語彙特徴と文脈特徴の双方に着目した文学作品の著者推定モデルを提案した．評価実験では，ベースラインに比べ精度が向上した．また実験結果から，語彙と文脈双方の特徴をとらえた著者推定を行えることが確認できた．これらのことから，提案手法が有効であることが示された．今後の課題として，より効果的な素性やアンサンブル手法の考案などが挙げられる．

また，今回は文学作品を段落ごとに分割し，段落単位での著者推定問題に取り組んだ．しかし，作品を章単位や作品単位で学習することによるグローバルな素性も考えられる．そういった素性を利用した著者推定問題にも今後取り組んでいきたい．

謝辞

本研究の一部は，科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した．

参考文献

[1] Mike Kestemont et al. 2018. Overview of the Author Identification Task at PAN-2018

[2] Douglas Bagnall. Author Identification using multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2015.

[3] Alberto Bartoli, Alex Dagri, Andrea De Lorenzo, Eric Medvet, and Fabiano Tarlao. An Author Verification Approach Based on Differential Features—Notebook for PAN at CLEF 2015.

[4] Houda Alberts, Author Clustering with the Aid of a Simple Distance Measure—Notebook for PAN at CLEF 2017

[5] Yasmany García-Mondeja, Daniel Castro-Castro, Vania Lavielle-Castro, and Rafael Muñoz. Discovering Author Groups using a B-compact graph-based Clustering—Notebook for PAN at CLEF 2017

[6] 馬場こづえ，藤井敦，石川徹也 小説テキストを対象としたジャンル推定と人物抽出 言語処理学会 2005

[7] 松浦司，金田康正 近代日本小説化8人による文章のn-gram分布を用いた著者判別 情報処理学会研究報告 2000

[8] 青空文庫 https://www. aozora. gr. jp/

[9] fastText Word vectors for 157 languages https://fasttext. cc/docs/en/crawl-vectors. html

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, andTomas Mikolov. 2017. Enriching word vectors withsubword information. Transactions of the Association forComputational Linguistics.