Vol.2015-CH-107 No.8
2015/8/9

データ解析に基づいた「花日記」の代作問題検証

孫　昊　　李　鍾賛　　金　明哲
同志社大学文化情報学研究科　同志社大学文化情報学部外国人研究員　同志社大学文化情報学学部

日本初のノーベル文学賞を受賞した川端康成にまつわる数多くの代作問題があり,その一つは「花日記」である.「花日記」は新潮社 1981 年版の川端全集第 20 巻に収録されているが,本作は当時川端康成を師事した主婦作家・中里恒子の代作という説がある.本研究は文章から抽出した文字・記号列の Bigram,タグの Bigram,文節パターン特徴量を基に,統合的分類アルゴリズムを用いて代作問題を検証した.

Ghostwriter verification of “Hananikki” based on data analysis

Hao Sun　　Jongchan LEE　　Mingzhe Jin
Graduate School of Culture and Information Science, Doshisha University
Visiting scholar of Faculty of Culture and Information Science, Doshisha University
Faculty of Culture Information Science, Doshisha University

Yasunari Kawabata, who is the first Japanese Nobel laureate in Literature has many ghost writers around him. One of his novels “Hananikki” is said to be written by his discipline Tsuneko Nakasato. In this research, several text measurements (bigram of words and punctuation mark, tags’ bigram, Phase Patterns) and the Integrated Classification Algorithm are used to discuss the authorship of “Hananikki”.

0. はじめに

網羅的な川端康成の全集として新潮社版の三十七巻本があり,そのうち二十三巻は小説である.本研究で取り扱う「花日記」(昭 13 年 4 月～昭 14 年 3 月)は第二十巻に収録され,川端康成の少女小説として知られている.しかし,川端康成全集補巻二「書簡来簡抄」に収録された中里恒子と川端康成の往復書簡に,「けふ少女之友買い,花日記にかかります.これは自分でも書いていてたのしみです.勿論虚構の人物ですけれどその人物に私の思っていることをみんなさせてゐるせいかもしれません.」のような記述があり,代作の疑惑が浮上していた.この書簡に基づき,「花日記」は中里恒子の書いた原稿に川端が手を入れて発表したという説がある（小野谷 2013）.

前述の書簡から川端康成と中里恒子二人ともこの「花日記」の創作に関わっていることが明らかになったが,関与の程度によって「花日記」の帰属が異なる.作品中の同性愛モチーフから関与度を探る試みはあったが,「花日記」の著者帰属問題を明らかにすることができなかった（大森 1991）.本研究では,計量文体分析の手法を導入し,文章から抽出した特徴データに基づいて「花日記」の代作問題を検証した.

1. 先行研究

日本語の計量文体分析研究の歴史は長く,文体計量分析の分野では様々な特徴量と分析手法が提案されてきた.特徴量は読点の打ち方,助詞の分布,タグの Bigram,文節パターンなどが有効である（Jin and Murakami 1993,金 2004,2013）.分析手法は多変量解析と機械学習の手法が多く用いられた.最近集団学習の考え方を取り入れた統合的判別法が提案された(金 2014）. 孫・金(2015)はこれらの特徴量と分析手法基づき,川端康成「山の音」の三島由紀夫による代作問題について分析を行った.

2. 用いたコーパス

文体は文章のジャンルと創作時期によって変る場合がある.本研究では, 出来る限りこのような影響要素を排除し,「花日記」に近い文章を用いてコーパスを構築した.

2.1 文体の統一化

今回の分析は小説だけを用いたので,小説のところどころ入っている日記,手紙や詩などのような小説と異なる文体の部分を削除した.登場人物の設定に合わせるため,文章中の会話文は故意に異なる文体で書かれる可能性があ

るので,本研究ではすべての会話文を削除した.

2.2 創作時期

同一著者が書いた文章でも創作時期によって文体が変る可能性がある.本研究では,検証対象作品「花日記」の創作時期（昭 13 年 4 月～昭 14 年 3 月）あたりの小説で,川端康成と中里恒子それぞれ 10 篇の小説を用いた.川端康成と中里恒子の作品リスト及び発表時期をそれぞれ表 1 と表 2 に示す.

表 1 用いた川端康成の作品リスト

発表時期	作品名
昭和 4 年 10 月	温泉宿
昭和 5 年 11 月	針と硝子と霧
昭和 7 年 2 月	抒情歌
昭和 8 年 7 月	禽獣
昭和 10 年 1 月	雪国
昭和 15 年 2 月	女の夢
昭和 15 年 3 月	ほくろの手紙
昭和 15 年 5 月	夜のサイコロ
昭和 15 年 6 月	燕の童女
昭和 15 年 7 月	夫唱婦和

表 2 用いた中里恒子の作品リスト

発表時期	作品名
昭和 12 年 6 月	自由画
昭和 13 年 1 月	ふみむすびと
昭和 13 年 7 月	西洋館
昭和 13 年 9 月	花火
昭和 13 年 11 月	樹下
昭和 14 年 6 月	森の中
昭和 14 年 9 月	野薔薇
昭和 14 年 9 月	乗合馬車
昭和 14 年 11 月	日光室
昭和 15 年 1 月	天国

3. 研究方法

本研究は記号論,形態論,構文論の視点に基づいて文章から 3 種類の特徴量を抽出した.その 3 種類の特徴量に複数の分類器による統合的判別法を用いて代作問題を検証した.先行研究では二人とも「花日記」の創作に関わっていることがわかり,本研究では章ごとに統合的判別法を応用して著者を推定した.

3.1 特徴量抽出

文章の計量分析のための特徴量は,出現頻度が多い,しかも文章の内容に依存しないものを用いるべきである.本研究は先行研究を踏まえて文字・記号の Bigram,形態素タグの Bigram,文節パターンの 3 種類を用いた.

(1) 文字・記号の Bigram

Bigram は 2 文字だけ続いた文字列のことである.文章の書き手識別に読点の打ち方（読点とその一つ前の文字・記号の組み合わせ）の有効性は報告されている（Jin and Murakami 1993）.読点の打ち方は,文字・記号の Bigram の一部に過ぎなく,本研究ではより多く情報を取るために,文字・記号の Bigram を用いた.

(2) タグの Bigram

本研究では,日本語形態素解析器 Mecab を用いて形態素解析を行った.例文:「学会で発表する.」の Mecab を用いた形態素解析の結果を次に示す.形態素の属性は階層化となっている.ここでは第三層までの情報を出している.

```
形態素　タグ
学会　　名詞,一般,*
で　　助詞,格助詞,一般
発表　　名詞,サ変接続,*
する　　動詞,自立,*
.　　記号,句点,*
```

書き手識別におけるタグの Bigram 特徴量の有効性が実証されており,ここで第一層のタグの Bigram を用いた（金 2004）.

(3) 文節パターン

構文論の視点から,四種類の文節パターンが提案されていた（金 2013）.本研究で用いる文節パターンは助詞・記号を除いた要素の第一層形態素タグの情報と助詞・記号の原型の組み合せである.例文:「学会で発表する.」を Cabocha で用いた係り受け解析の結果を次に示す.二つの「*」の間に入っている部分が一つの文節である.この例文では,第一文節の文節パーンは「名詞_で」,第二文節の文節パターンは「名詞_動詞_.」となる.

```
* 0 1D 0/1 0.000000
学会　　名詞,一般,*,*,*,*,学会,
で　　助詞,格助詞,一般,*,*,*,で,
* 1 -1D 1/1 0.000000
発表　　名詞,サ変接続,*,*,*,*,発表,
する　　動詞,自立,*,*,サ変・スル,基本形,する,
.　　記号,句点,*,*,*,*,.,.,.
```

3.2 判別方法

本研究ではランダムフォレスト(RF:Randomforest), サポートベクターマシン(SVM:Support vector machine)を用いる.川端康成と中里恒子の文章から抽出した各特徴量を学習データとし,「花日記」の各章から抽出した各特徴量をテストデータとして検証を行った.各特徴量・分類器における予測分類結果の確率平均値を取って著者の帰属を推定した.

4. 結果分析

文字・記号列の Bigram,タグの Bigram,文節パターンによる判別結果をそれぞれ表 3,表 4,表 5 に示す.文字・記号列の場合,「花日記」の 12 章は全て川端康成に属する.タグの Bigram の場合,第 2,9,10 章は中里恒子作で,他の章は川端康成作という結果が得られた.文節パターンの場合,第 2,10 章は中里恒子作で,他の章は川端康成作と判別された.

表 3 文字の Bigram による統合的判別

花日記(各章)	文字・記号の Bigram				川端平均	中里平均	判別
	SVM		RF				
	川端	中里	川端	中里			
第 1 章	0.561	0.439	0.594	0.406	0.577	0.423	川端康成
第 2 章	0.519	0.481	0.564	0.436	0.541	0.459	川端康成
第 3 章	0.539	0.461	0.562	0.438	0.551	0.449	川端康成
第 4 章	0.506	0.494	0.548	0.452	0.527	0.473	川端康成
第 5 章	0.526	0.474	0.622	0.378	0.574	0.426	川端康成
第 6 章	0.503	0.497	0.636	0.364	0.569	0.431	川端康成
第 7 章	0.536	0.464	0.550	0.450	0.543	0.457	川端康成
第 8 章	0.538	0.462	0.582	0.418	0.560	0.440	川端康成
第 9 章	0.524	0.476	0.844	0.456	0.534	0.466	川端康成
第 10 章	0.514	0.486	0.646	0.354	0.580	0.420	川端康成
第 11 章	0.498	0.502	0.616	0.384	0.557	0.443	川端康成
第 12 章	0.549	0.451	0.578	0.422	0.564	0.436	川端康成

表 4.タグの Bigram による統合的判別

花日記(各章)	タグの Bigram				川端平均	中里平均	判別
	SVM		RF				
	川端	中里	川端	中里			
第 1 章	0.501	0.499	0.438	0.562	0.469	0.531	中里恒子
第 2 章	0.632	0.368	0.520	0.480	0.576	0.424	川端康成
第 3 章	0.387	0.613	0.402	0.598	0.394	0.606	中里恒子
第 4 章	0.318	0.682	0.500	0.500	0.409	0.591	中里恒子
第 5 章	0.292	0.708	0.470	0.530	0.381	0.619	中里恒子
第 6 章	0.278	0.722	0.428	0.572	0.353	0.647	中里恒子
第 7 章	0.378	0.622	0.458	0.542	0.418	0.582	中里恒子
第 8 章	0.440	0.560	0.548	0.452	0.494	0.506	中里恒子
第 9 章	0.486	0.514	0.564	0.436	0.525	0.475	川端康成
第 10 章	0.773	0.227	0.554	0.446	0.663	0.337	川端康成
第 11 章	0.335	0.665	0.528	0.472	0.431	0.569	中里恒子
第 12 章	0.323	0.677	0.532	0.468	0.427	0.573	中里恒子

表 5. 文節パターンによる統合的判別

花日記(各章)	文節パターン				川端平均	中里平均	判別
	SVM		RF				
	川端	中里	川端	中里			
第 1 章	0.501	0.499	0.472	0.528	0.486	0.514	中里恒子
第 2 章	0.632	0.368	0.508	0.492	0.570	0.430	川端康成
第 3 章	0.387	0.613	0.482	0.518	0.434	0.566	中里恒子
第 4 章	0.318	0.682	0.484	0.516	0.401	0.599	中里恒子
第 5 章	0.292	0.708	0.612	0.388	0.452	0.548	中里恒子
第 6 章	0.278	0.722	0.482	0.518	0.380	0.620	中里恒子
第 7 章	0.378	0.622	0.474	0.526	0.426	0.574	中里恒子
第 8 章	0.440	0.560	0.554	0.446	0.497	0.503	中里恒子
第 9 章	0.486	0.514	0.448	0.552	0.467	0.533	中里恒子
第 10 章	0.773	0.227	0.582	0.418	0.677	0.323	川端康成
第 11 章	0.335	0.665	0.584	0.416	0.459	0.541	中里恒子
第 12 章	0.323	0.677	0.564	0.436	0.443	0.557	中里恒子

表 6 の全ての特徴量を用いた統合結果から見ると,第 1,2,8,9,10 章は川端康成作となり,第 3,4,5,6,7,11,12 章は中里恒子作となる.

表 6. 全ての特徴量による統合的判別

花日記	川端平均	中里平均	判定
第 1 章	0.511	0.489	川端康成
第 2 章	0.562	0.438	川端康成
第 3 章	0.460	0.540	中里恒子
第 4 章	0.446	0.554	中里恒子
第 5 章	0.469	0.531	中里恒子
第 6 章	0.434	0.566	中里恒子
第 7 章	0.462	0.538	中里恒子
第 8 章	0.517	0.483	川端康成
第 9 章	0.509	0.491	川端康成
第 10 章	0.640	0.360	川端康成
第 11 章	0.482	0.518	中里恒子
第 12 章	0.478	0.522	中里恒子

5. 結論

直しやすいと思われる文字・タグの Bigram 特徴量では,「花日記」の 12 章は全て川端康成と判別されたので,確かに川端は一通り手を入れていたことがわかった.

直しにくいと思われるタグの Bigram 特徴量と文節パターン特徴量両方とも川端康成作となる第 2 章と第 10 章に関しては,川端康成はしっかり書き直したと思われる.

タグの Bigram 特徴量と文節パターン特徴量だけでなく,全ての特徴量を用いた判別結果でも,中里恒子の書いたと判別された章が多い.

6. 課題

本研究は,小規模のコーパス及び二つの分類器に基づいた検証だけである.今後コーパスと分類器を増して検証することが必要である.

参考文献

[1] 小谷野敦(2013).「川端康成伝－双面の人」,中央公論新社.
[2] 大森郁之助 (1991) .「乙女の港」・その地位の検証 ： lesbianism の視点ほか,または,八木洋子頌. 札幌大学女子短期大学記要 17,A1-A18.
[3] Jin, M. and Murakami, M. (1993). Author’s Features Writing Styles as Seen Through Their Use of Commas, Behaviormetrica,20,1,63-76.
[4] 金明哲(2004).品詞のマルコフ遷移の情報を用いた書き手の同定,日本行動計量学会第 32 回全国大会講演抄録集,384-385.
[5] 金明哲(2013).文節パターンに基づいた文章の書き手の識別,行動計量学,40,1,17-28.
[6] 金明哲(2014).統合的分類アルゴリズムを用いた文章の書き手の識別.行動計量学,第 41 巻,第 1 号, 35-46.
[7] 孫昊,金明哲(2015).川端康成「山の音」の代筆疑惑検証―計量文体学の観点から―,言語処理学会第 21 回年次大会発表論文集.