役者評判記からの人物表現抽出手法の提案

永井 規善
立命館大学 情報理工学研究科

前田 亮
立命館大学 情報理工学部

木村 文則
立命館大学 衣笠総合研究機構

赤間 亮
立命館大学 文学部

本論文では，役者に関する評判について書かれた江戸時代の書籍である役者評判記から，人名や別名などの表現を抽出する手法を提案する．本手法では，文字単位に機械学習を行う固有表現抽出手法を応用して人名や別名などの表現を抽出した．役者評判記本文のテキストデータと人手で付与された注釈データを用いて実験を行った結果，交差検定により F 値で約 0.91 という高い結果が得られた．

Personal Name Extraction from Japanese Historical Document of Actor Reviews, “Yakusha-Hyoban-Ki”

Noriyoshi Nagai[1] Akira Maeda[2] Fuminori Kimura[3] Ryo Akama[4]
[1]Graduate School of Information Science and Engineering, Ritsumeikan University
[2]College of Information Science and Engineering, Ritsumeikan University
[3]Kinugasa Research Organization, Ritsumeikan University
[4]College of Letters, Ritsumeikan University

In this paper, we propose a method for extracting persons’ real names and aliases from Japanese historical document of actor reviews, “Yakusha-Hyoban-Ki”. In this method, we extract personal names and aliases by applying a named entity extraction technique using machine learning based on character units. Experimental results showed that our proposed method were able to extract personal names and aliases from “Yakusha-Hyoban-Ki” with approximately 0.91 in F-measure by the cross-validation.

１．はじめに

近年，古典史料の電子テキスト化が進み，それらを集積したデータベースの構築や古典史料ディジタルアーカイブが Web 上で公開され，今後より一層古典史料を解析，活用する研究が進むと考えられる．古典史料の分析は，調査対象となる事物の出現回数を網羅的に集計した結果を用いた分析が一般的である．そのため，調査対象外となる事物について集計されることは少なく，比較調査を行うといったことは容易ではない．また，これらの研究の多くは手作業によって行われており，労力を要する．コンピュータを用いて電子テキスト化や分析を補助することで，調査対象の事物以外の幅広い調査や，研究の可能性を広げることが可能である．

本研究では江戸時代に刊行された役者のレビューである役者評判記を用いる．役者評判記はこれまで，役者評判記研究会 98 によってほぼ手作業で電子テキスト化や注釈付け，索引の作成などが行われてきたが，一部自動化されている部分がある．例えば，注釈付けには Web 上で行える注釈付け・閲覧システムを作成し，利用している．このシステムでは，役者目録から作成した人名索引とのマッチングによって，本文テキスト中で目録と完全に文字列が一致している箇所が自動で注釈候補として示される．しかし，略称，俳名，屋号などの別名に関しては現状では網羅している索引がなく，網羅的に抽出することは人手でも困難である．また，誤って注釈付けの必要がない文字列が注釈候補として示されることもある．そこで本研究では，役者評判記の分析に役立てるため，人物の実名や別名などの固有表現，すなわち人物表現と，人物表現の前後の単語や文字出現パターンを機械学習することで，役者評判記に含まれる人物表現を網羅的に抽出する手法を提案する．

(c) Information Processing Society of Japan

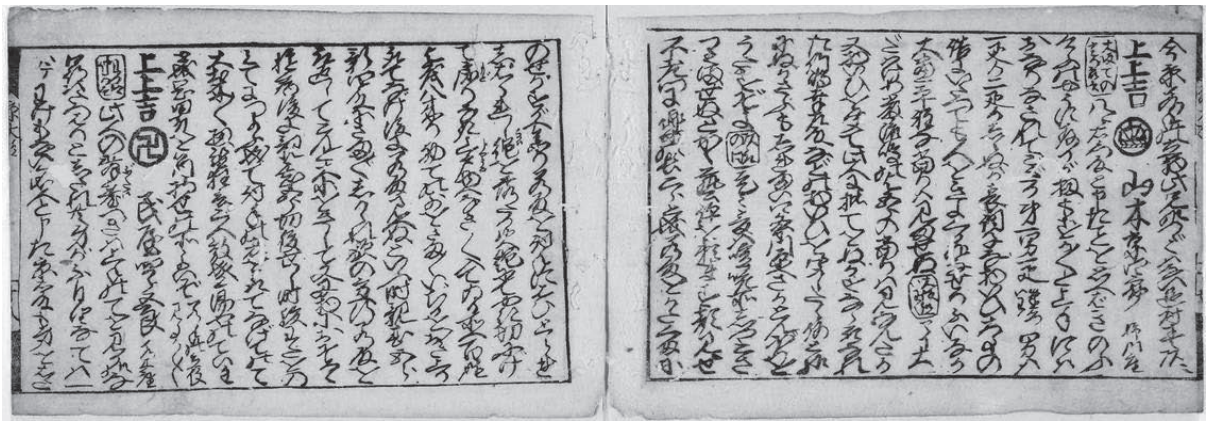

図 1　役者評判記『役者大極舞』
立命館大学ＡＲＣ所蔵(arcBK04-0033・18 頁)

２．役者評判記

役者評判記は江戸時代に刊行された書物で，歌舞伎役者の評判に関して書かれたものである．長期間定期的に刊行されたという特徴から，役者評判記は演劇史においても貴重な史料となっている．図 1 は 1739 年の『役者大極舞』の一部である．

役者評判記は，『吉』と『上』などを用いた位付け，役者名，座元，そして評者らの問答による評価文，という形式で記述されている．図 2 にその例を示す．

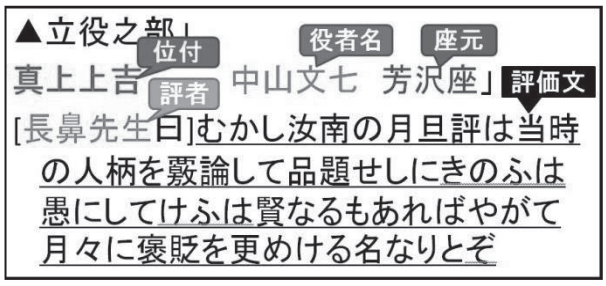

図 2　役者評判記『役者一陽来』の一部

歌舞伎役者の人物表現は，役者名以外に屋号や俳名などがある．屋号はその歌舞伎役者の出身地や副業などに由来した呼ばれ方である．また，俳名は俳句を詠む際に用いる別名で，多くの歌舞伎役者が俳名を持っており，先代から芸名を襲名する際に俳名も譲り受けることもあった．このように，役者評判記には様々な人物表現が含まれる．

現在，役者評判記研究会 98 が翻刻本文の作成やデータの管理，注釈付けなどを行っている．本研究では当研究会からの許諾と協力を得て，第三期役者評判記テキストデータとその注釈データを使用させてもらった．

３．関連研究

井坪ら[1]は平安時代から鎌倉時代にかけて記された『兵範記』，『玉葉』，『吾妻鏡』を研究対象とし，それぞれの人物関係の時間的変化や地名を通した人物関係を推定し，可視化する手法を提案した．人名抽出にはそれぞれの史料から人手で作成された人名索引を利用した．これらの索引には登場人物の実名だけでなく，登場した日付や同一人物を示す表現なども登録されており，別の表現で表記されている場合でも特定が可能で，より正確に登場人物の出現頻度を得ることが可能である．本研究の対象である役者評判記は，役者評判記研究会 98 にて作成された人名索引が利用可能であるが，俳名や屋号などの別名の表現は網羅されていないため，本研究では人物表現を推定する手法を提案する．

山田ら[2]は機械学習アルゴリズム，Support Vector Machines（SVM）を用いて現代日本語の固有表現抽出を提案した．SVM は教師あり学習と線形入力素子を利用する二値線形分類器である．単語ごとに解析を行い，単語自身，品詞分類，文字種，およびこれらの組み合わせを素性として使用している．実験には CRL 固有表現データを使用していて，これは毎日新聞 95 年度版 1,174 記事，約 11,000 文に対して固有表現が付与されているものである，SVM を用いた抽出実験にて F 値で 0.83 という精度を得ている．

浅原ら[3]は山田ら[2]の研究に対し，テキストを文字単位に分割し，文字ごとに解析を行っている．これにより固有表現となる語の境界をより正確に判定し，SVM を用いた抽出実験にて F 値で約 0.87 という精度を得ている．

吉村ら[4]は山田ら[2]や浅原ら[3]による固有表現抽出手法を参考に，文字の出現頻度と文字列の出現確率を利用し，漢文体などの形態素解析がで

(c) Information Processing Society of Japan

きない古文テキストに対しての人物表現の抽出手法を提案した．本研究では，吉村らの手法を基にし，単語分割情報に中古和文 UniDic[5]を利用することで，人物表現抽出精度を向上させる．

4．古文形態素解析器の利用

小木曽ら[5]は和文系の資料を対象とした形態素解析辞書である中古和文 UniDic を開発した．中古和文 UniDic は『源氏物語』を始めとする中古和文を対象とする形態素解析辞書で，主に MeCab[6]の辞書として用いる．小木曽らによると，中古和文 UniDic で『源氏物語』,『大和物語』,『土佐日記』,『紫式部日記』を形態素解析した場合の精度は 95%を超えており，中古和文を高い精度で解析することが可能である．

古文に対して形態素解析を行うことができる他の形態素解析辞書として，小木曽らが開発した近代文語 UniDic[7]が挙げられる．これは近代の文語論説文を対象とした形態素解析辞書である．

近代文語文と中古和文では，役者評判記の刊行された江戸時代に年代が近いのは近代文語である．しかし，明治時代に文体と口語体を一致させる，言文一致運動というものが行われ，これにより平安時代から江戸時代まで続いた和文から近代文語文へと変化していったという経緯がある．本論文で扱う役者評判記は江戸時代に刊行されたもので，当時の文体は平安時代の和文の語法や用語，文体を真似して用いられたものである．これらのことから本研究では，より取り扱っている文体が近い，中古和文 UniDic を利用して形態素解析を行い，単語分割の推定情報として SVM の素性に利用する．

5．提案手法

提案手法では本文中の各文字を，人物表現に含まれる文字かどうかを分類することで，人物表現を抽出する．提案手法の概要を図 3 に示す．まず，中古和文 UniDic を利用し，役者評判記テキストデータの単語分割情報を得る．その後，それぞれの文字に単語分割情報や前後の文字情報などの素性の付与を行う．学習データは，役者評判記の既存の注釈データから人物表現の正解データを取り出すことで作成する．次に SVM を用い学習データからモデルを作成し，テキストデータの各文字を分類する．そして分類したラベルから人物表現の推定，抽出を行う．

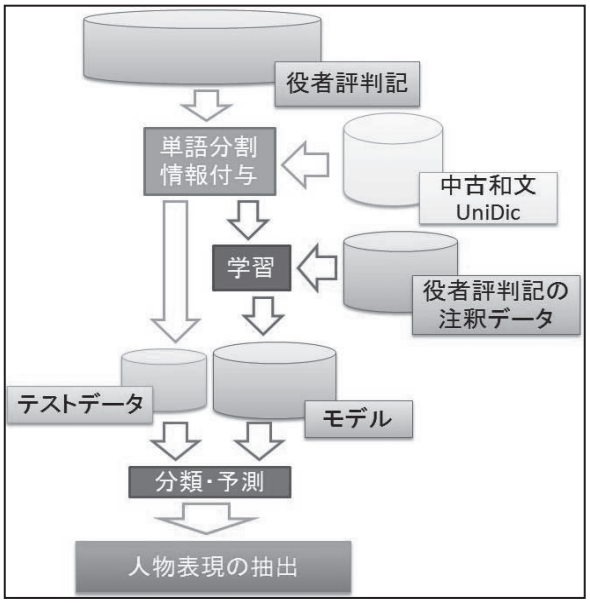

図 3　提案手法の概要図

5.1 単語分割情報

吉村らは形態素解析が困難である古文テキストを対象としていたため，文字 N グラムの出現確率を利用した単語分割[8]を行った．今回我々は中古和文 UniDic を用いて役者評判記テキストを形態素解析し，各形態素を単語と見なして単語分割情報に用いる．単語分割情報を各文字に付与する際には浅原ら[3]が用いている Start/End（SE）タグを利用する．単語分割結果の先頭の文字に“B”のタグ，末尾の文字に“E”のタグ，内部の文字には“I”のタグ，1 文字なら“S”のタグを付与する．この SE タグを付与した分割結果を学習の素性として利用する．役者評判記の一部を SE タグに分類した例を表 1 に示す．「七代めの団十郎と」を中古和文 UniDic による形態素解析で分割すると「七｜代｜め｜の｜団十郎｜と」となる．この中の「団」は「団十郎」の先頭の文字なので「B-団十郎」,「十」は内部の文字なので「I-団十郎」,「郎」は末尾の文字なので「E-団十郎」となる．

表 1　SE タグに分類した例

文字	SEタグを付与した分割結果
七	S-七
代	S-代
め	S-め
の	S-の
団	B-団十郎
十	I-団十郎
郎	E-団十郎
と	S-と

(c) Information Processing Society of Japan

5.2 人物表現の抽出

提案手法では文字ごとに処理を行い，分類するため，最後に文字の並びを人物表現になるよう統合する．人物表現に統合する例を図４に示す．

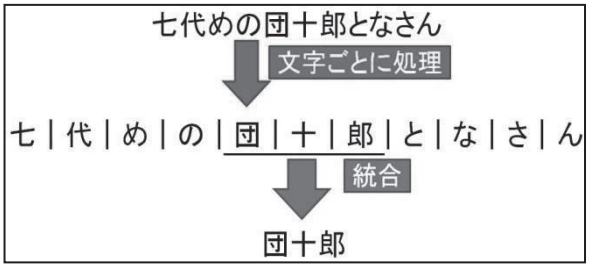

図 4　人物表現を統合した例

次に，各文字を人物表現に統合する手法について説明する．分割された各文字の集合を人物表現として表現するため，IOB2 タグ集合を用いる．これは，人物表現の先頭の文字に“B”のタグ，人物表現の先頭以外の文字に“I”のタグ，人物表現以外の文字に“O”のタグを付与するものである．入力文の各文字を IOB2 タグに分類する規則を学習することで，人物表現の抽出を行うことができる．役者評判記の一部を IOB2 タグに分類した例を表２に示す．表２では，人物表現の先頭の文字「団」が B，人物表現の先頭以外の文字「十」，「郎」が I，それ以外の文字が O となる．この表のように分類された場合，B，I，I と並んでいる「団十郎」を人物表現として統合することができる．

表 2　IOB2 タグに分類した例

文字	IOB2タグ
七	O
代	O
め	O
の	O
団	B
十	I
郎	I
と	O

5.3 使用する素性

学習データに使用する素性について記述する．文頭から i 番目の文字に関する素性は，i-2 番目から i+2 番目までの各文字と単語分割情報と文字の位置，i-2 番目と i-1 番目の IOB2 タグである．使用する素性を表３に示す．網掛けをしてある部分が i 番目の文字に付与されている素性である．

表 3　使用する素性の例

位置	文字	SEタグを付与した分割結果	IOB2タグ
i-2	め	S-め	O
i-1	の	S-の	O
i	団	B-団十郎	B
i+1	十	I-団十郎	I
i+2	郎	E-団十郎	I

IOB2 タグは学習時には既知であるが，解析時には未知であるため，それぞれの位置で推定した IOB2 タグをその次の文字の素性としても利用する．

これらの素性を SVM による機械学習の素性として利用する．しかし SVM は二値線形分類器であるため，提案手法のように，分類するクラスが3以上ある場合には多値分類に拡張する必要がある．本手法では吉村ら[4]に倣い one-versus-rest 法を用いた多値分類を行う．one-versus-rest 法は k 個のクラスに対し，あるクラスかそれ以外かを分類する二値分類器を k 個構築する手法である．

６．評価実験

前節で述べた提案手法により役者評判記テキストデータから人物表現を抽出する実験を行った．正解データには人手で付与された注釈データのうち，人物表現に関連するものを利用した．

6.1 実験データ

実験データとして役者評判記の古文テキスト 231 冊分のデータと，これらに付けられた注釈データを利用した．このデータ全体の文字数と人物表現の出現回数を表４に示す．

表 4　役者評判記の文字数と人物表現の出現回数

テキストの文字数	人物表現の出現回数
2,004,572	61,145

これらの役者評判記の注釈は，すべてが人手で付けられているものではない．しかし，人手で作成した人名索引との文字列マッチングにより，大半の人物表現が自動的に注釈付けされ，分類が付与されている．分類には「演目名」，「作品名」，「人名」，「役者」，「作者」，「興行関係者」，「役名」，「事項」，「地名」，「劇場」，「建物名」などがあり，本研究ではこれらのうち「人名」，「役者」，「作者」，「興行関係者」，「役名」の分類が付けられた単語

(c) Information Processing Society of Japan

を人物表現の正解データとして機械学習に利用した.

6.2 評価方法

提案手法による抽出結果と人手で付けられた注釈を比較することにより，正解を判定する．実験による抽出数から適合率，人物表現のタグが付けられた注釈から再現率を算出し，それらの調和平均である F 値を算出する．それぞれの計算式を以下に示す.

$$\text{適合率} = \frac{\text{抽出結果に含まれる正解の数}}{\text{抽出した人名表現の数}}$$

$$\text{再現率} = \frac{\text{抽出結果に含まれる正解の数}}{\text{正解の人物表現の総数}}$$

$$\text{F 値} = \frac{2\times\text{適合率}\times\text{再現率}}{\text{適合率}+\text{再現率}}$$

評価には役者評判記テキストデータを 5 等分し，学習 4，テスト 1 の比率で交差検定を行いそれらの適合率，再現値，F 値の平均をそれぞれ用いる．また，中古和文 UniDic による単語分割情報を利用する場合，どの程度精度に影響するのかを確認するため，単語分割情報を利用する場合としない場合で比較を行う．単語分割情報を利用しない場合を「比較対象」，利用した場合を「提案手法」とし，表 5，表 6 にそれぞれの場合での実験結果を示す．また，図 5 に適合率，再現率，F 値それぞれの値を比較するグラフを示す.

表 5　比較対象における実験結果

適合率	再現率	F値
0.8635	0.7691	0.8136

表 6　提案手法における実験結果

適合率	再現率	F値
0.9289	0.9100	0.9193

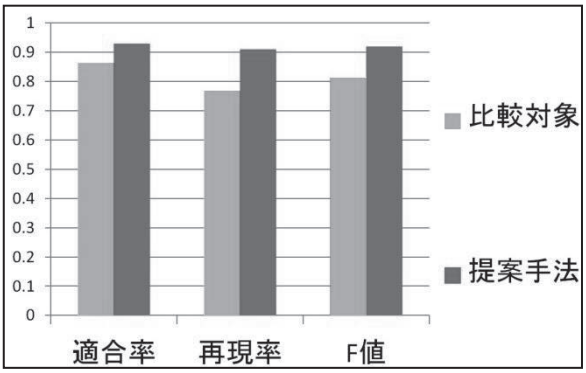

図　5　比較対象と提案手法における抽出実験結果

7．考察

表 5，6 から単語分割結果の情報を利用することで，F 値で約 10 パーセント程度の精度が向上することがわかった.

提案手法による人物表現の抽出結果に，本来正解と判定されるべきであるが，不正解と判定されてしまっていた人物表現がいくつか含まれていることがわかった．まず，別名として正しく抽出できていた例を図 6 に示す．下線と赤文字が付けられた部分がうまく抽出できた人物表現である.

*大竹や*のむすこもいつしかこのやうに成人～

図　6　「大竹や」の抽出例

「大竹や」は正解データにない人物表現である．類似する人物表現として「大竹屋」が正解データに存在する．「大竹屋」は「沢村淀五郎」という役者の屋号である．屋号とは一門，一家の特徴を基に家に付けられる称号のことで，江戸時代当時では呼び名として使うことがあった．「大竹や」は役者評判記の本文のうち，評者による評価文にて使われている表現で，「屋」を書く際に「や」と書いたものと推測できる．この結果から，「大竹屋」での完全一致では抽出できない人物表現が提案手法によって抽出できていることがわかる.

次に，正解として判定されなかったが，正しく抽出されていた例を図 7 に示す.

かくせし大鼓をとり出し中居*おさく*をいひ名づけの女房と知り～

図　7　役名「おさく」の抽出例

中居である「おさく」は歌舞伎に登場する役の一つである．「おさく」には注釈が付けられていないため，人物表現として抽出されたものの，正解として判定されなかった．注釈は一般的に理解に役立てるための補足情報であるので，このように役名であることが文脈から明らかな場合には付けられないことがある．このことは，コンピュ

(c) Information Processing Society of Japan

ータ処理で古典史料から人物表現を網羅することの重要性を示唆している．また，表 7 に「おさく」に付けられた素性の例を示す．

表 7　役名「おさく」に付けられた素性の例

位置	文字	SEタグを付与した分割結果	IOB2タグ
i-2	中	S-中	O
i-1	居	S-居	O
i	お	B-お	B
i+1	さ	I-さ	I
i+2	く	E-く	I

「おさく」にはこのような素性が付与されていた．SE タグを付与した分割結果から，「おさく」が一つの単語として出力されていることがわかる．しかし，中古和文 UniDic でこの文章を形態素解析した結果によると，「おさく」は「押す」という動詞の活用形という誤った解析結果であった．このことから，役者評判記の擬古文に対して中古和文UniDicでの形態素解析を行う場合と，文体や語彙の相違により必ずしも正しい結果が得られるとは限らず，これにより人物表現を正しく抽出できない可能性が考えられる．したがって，本論文で使用した中古和文 UniDic による単語分割以外の手法の検討も今後必要であると考えられる．

8．おわりに

本論文では機械学習を用いて役者評判記から人物表現を抽出する手法を提案した．中古和文 UniDic での形態素解析結果を分割情報として利用し，役者評判記から人物表現の抽出を行った．結果として F 値で 0.91 という高い精度で人物表現の抽出ができた．

本研究の今後の課題として，まず人物表現の抽出精度をさらに向上させため，まず，単語分割手法について再検討する．たとえば歌舞伎辞典の語彙などを中古和文 UniDic に追加することで，単語分割の精度をより向上させることができるのではないかと推測する．

また，本手法の応用として，歌舞伎における役者名の変遷を抽出する手法を検討している．歌舞伎役者は子役，若手，中堅，大御所，そして引退時，複数の名前を持つため，これらを役者評判記の文章から推測できる手法を検討している．

さらに，評判の記述と位付けから歌舞伎役者に対する評価を分析する手法を検討している．

謝辞

役者評判記のテキストデータおよび注釈データの使用を許可していただいた役者評判記研究会 98 の皆様に深く感謝申し上げます．

参考文献

1)　井坪将，木村文則，前田亮：古典史料からの相対的な人物関係の時間的変化の推定と可視化，人文科学とコンピュータシンポジウム論文集，pp.29-36（2011）．

2)　山田寛康，工藤拓，松本裕治： Support Vector Machines を用いた日本語固有表現抽出， 情報処理学会論文誌，Vol.43，No.1，pp.44-53（2002）．

3)　浅原正幸，松本裕治： 日本語固有表現抽出におけるわかち書き問題の解決，情報処理学会論文誌，Vol.45，No.5，pp.1442-1450（2004）．

4)　吉村衛，木村文則，前田亮：古文テキストからの人物表現抽出，人文科学とコンピュータシンポジウム論文集，pp.97-102（2013）．

5)　小木曽智信，小椋秀樹，田中牧郎，近藤明日子，伝康晴：中古和文を対象とした形態素解析辞書の開発， 情報処理学会研究報告 人文科学とコンピュータ， Vol.2010-CH-85（No.4）pp.1-8（2010）．

6)　Taku Kudo，Kaoru Yamamoto，Yuji Matsumoto：Applying Conditional Random Fields to Japanese Morphological Analysis，Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing（EMNLP-2004），pp.230-237（2004）．

7)　小木曽智信，小椋秀樹，近藤明日子: 近代文語文を対象とした形態素解析辞書の開発，言語処理学会第 14 回年次大会発表論文集，pp.225-228（2008）．

8)　吉村衛，木村文則，前田亮：古文テキスト解析のための文字 N グラムの出現確率を利用した単語分割，人文科学とコンピュータシンポジウム論文集，pp.261-268（2011）．

(c) Information Processing Society of Japan