役者評判記を用いた役者情報の抽出

川端　恵大（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）
赤間　亮（立命館大学　文学部）

概要：古典資料の索引作成は手作業に頼ることが多い．自然言語処理技術を用いて自動で索引に用いられる単語を抽出できれば，歴史・文化研究を行う専門家の支援が可能になると考えられる．本研究では，役者評判記と呼ばれる江戸時代から明治初期まで発行された古典資料から歌舞伎役者に関する情報を抽出する実験を行い，役者情報抽出の精度を求めた．
キーワード：役者評判記，古典文書，固有表現抽出

Named entity recognition for Yakusha hyobanki

Keita Kawabata (Graduate School of Information Science and Engineering, Ritsumeikan University)
Akira Maeda (College of Information Science and Engineering, Ritsumeikan University)
Ryo Akama (College of Letters, Ritsumeikan University)

Abstract: Index creation of historical documents often relies on manual work. If natural language processing technology can be used to automatically extract words for indexing, it will be possible to support specialists in historical and cultural research. In this study, we conducted an experiment to extract information about Kabuki actors from historical documents called Yakusha hyobanki, which were published from the Edo period to the early Meiji period, and determined the accuracy of the extraction of actor information.
Keywords: Yakusha hyobanki, Historical document, Named Entity Recognition

1.　まえがき

古典資料のデジタル化は OCR（光学文字認識）術の発展[1]により，人による手作業から機械による自動化が主流となりつつある．しかし，古典資料内の単語の索引を作成する場合，いまだに手作業に頼る部分が大きい．この索引の作成作業を自動で行うことができれば，歴史・文化研究を行う専門家の支援が可能となり，研究を効率的に行うことが可能となる．

本研究では，役者評判記と呼ばれる江戸時代から明治初期まで発行された歌舞伎役者の批評書から深層学習モデルである BiLSTM を用いて役者情報の抽出を行う実験を行なった．図 1 に本研究で使用する役者評判記の原文の一部を示す．

図 1　役者評判記『役者多名卸（江戸）』
出典：役者評判記『役者多名卸』3 項
立命館 ARC(BK04-0184)

2.　役者評判記

役者評判記は，江戸時代から明治時代初期まで京，大阪，江戸で毎年刊行された歌舞伎役者の芸評書である．歌舞伎役者を役柄ごとに分類し，複数の評価者が評価を行っている．大きく分けて「役者の紹介」，「挿絵」，「役者の評価，批評」という構造になっている．本研究で使用する役者評判記は，役者評判記研究会 98[2]によってデジタルテキスト化されており，挿絵に記載されている文字も含めてデジタル化されている．

図 2 は役者評判記『役者多名卸（江戸）』のテキストデータの一部である． デジタルテキスト化される際には, 図 2 のように構造化されている. 吹き出しや文字の色は著者によって付加したものである．この例では，役種は「立役之部」である．「立役之部」とは役種の立役を指しており，男役を意味する．歌舞伎役者の演技の評価は位付と呼ばれる評価方法が用いられており「極」，「上」，「吉」といった文字を用いて評価を行う．役者名は「市川海老蔵」であり, 座本と呼ばれる歌舞伎の興行を取り仕切る組織の情報が書かれおり, 歌舞伎役者がどの座元に所属しているかがわかる. そして,「兄大臣十介」による評価が書かれていることがわかる. 評価文の中にも役者名や座元の情報を表す語が多く出現する．

図 2　役者評判記『役者多名卸（江戸）』のテキストデータの一部

3.　関連研究

本研究では, デジタル化された役者評判記から役者情報を抽出するための情報抽出手法として固有表現抽出を使用する. 固有表現抽出とは，テキスト内に存在する「人名」，「組織」といった固有名詞や「時刻」，「数値」といった数値情報のような特定の情報を抽出する情報抽出技術の呼称である. 固有表現抽出に用いられる手法として SVM（Support Vector Machine）や CRF（Conditional Random Fields）が存在する．近年，深層学習に注目が集まるにつれ LSTM（Long Short-Term Memory）をベースとした研究が多く報告されている. LSTM はニューラルネットワークの一種である RNN（Recurrent Neural Network）における文脈や単語の長期記憶の問題を解決したモデルである．

Huang ら[3]が提案した BiLSTM（Bidirectional Long Short-Term Memory）-CRF は，単語の分散表現を入力とし CRF 層で単語間の依存関係を考慮したラベルの予測をする. 固有表現抽出を行うためには, 抽出対象となる単語にラベルを付与し分類する必要がある. Huang らが提案した BiLSTM-CRF モデルは，CoNLL2003[4]やその他のデータセットでも高い抽出精度を示している．

Lample ら[5]は，Huang らの提案手法を応用し，単語の文字の分散表現を入力とし BiLSTM-CRF モデルに学習させることで, 単語の分散表現を入力としたモデルよりも高い精度を示した．

矢野ら[6]は, 医療文書を対象に病名, 疾患名の識別と患者が患っている疾患の所見の分類タスクに BiLSTM-CRF モデルを使用し実験を行った. 入力に使用されたデータセットは文字をベースに疾患の所見を示すラベルがアノテーションされている．また，その他に ICD-10 コードと呼ばれる疾患の種類を表す数字と, 表層文字が「ひらがな」や「漢字」といった文字情報を持っていることをあらわす文字タイプ情報を加えて, 複数の素性を用いて比較実験を行っている. この結果, CRF のみを用いた実験より高い精度を示した．

白井ら[7]は，芳賀矢一著『日本人名辞典』を対象として BiLSTM+CRF モデルを用いて固有表現抽出による人物の関係性を抽出する研究を行なった. その結果から，特定のパターンで記述されている関係に関して高い抽出精度を得られることを示した．

4.　データセットとその処理

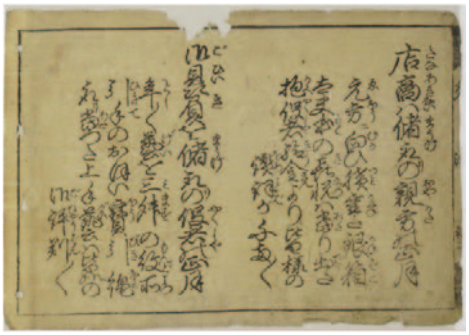
デジタル化された役者評判記には古典文書特有の表記を表現するために複数の記号を用いている. 通常の文字より小さく表現されている文字の後には「@」, 原文ではひらがなの「く」のように表現されている文字は「くの字点」と呼ばれるもので, デジタルテキスト上では「／＼」や「／”＼」と表現されている．また，図 2 の位付で出現する「極」を囲む記号「＜＞」も原文の表現方法の代わりに用いられる記号の 1 つである．その例を図 3 に示す．これらの記号は形態素解析の妨げとなる可能性が高いと考え除去する．

役者の位付と役者の名前の間に, 歌舞伎役者の家紋が描かれている場合がある. デジタルテキストでは「（紋）」と表現されている．こちらも抽出の妨げになると考え除去する. 図 4 に「（紋）」が記述されている例を示す．ストップワードとして削除する記号や文字を表 1 にまとめる．

表 1 で示したストップワードとして除去する記号とは逆に, 役者情報を識別するために有用であると考えられる記号も存在する. 例えば「▲」や「[ ]」である. 「▲」は「▲立役之部」のように役種を表し, 「[ ]」は「[兄大臣十介曰]」のように評価者を表す. このような記号情報は役者情報抽出の助けとなると考え, モデルの学習の際に使用する.

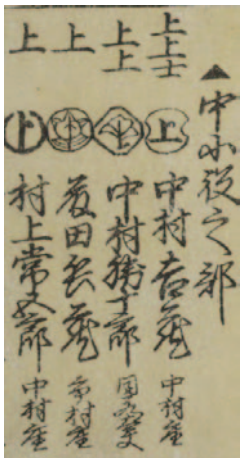
表 1 除去する記号

記号	意味
@	小さい文字の補助
／＼	くの字点
／″＼	くの字点
(紋)	家紋
＜＞	文字の強調



店商は儲取り@の親方の正月
元方二@向ひ積重た銀箱
太夫本の喜悦は当り出た
抱役者給金より皆様の讃詞が千両／＼
御贔屓を儲取り@の役者の正月
年／＼芸を三升の紋所
引手のおほい宝引縄
取り@当つた上手芸は皆様の御評判／＼

図 3 原文とデジタルテキストの比較
出典：役者評判記『役者多名卸』3 項の一部
立命館 ARC(BK04-0184)



▲中小役之部
上上半吉(紋)中村吉蔵　中村座
上上(紋)中村勝十郎　同若大夫
上(紋)藤田兵蔵　市村座
上(紋)村上常五郎　中村座

図 4 「（紋）」が記述される例
出典：役者評判記『役者多名卸』30 項の一部
立命館 ARC(BK04-0184)

本研究で抽出する役者情報は「役者名」「座本」「役種」「位付」「評価者」である. 表 2 に抽出項目とそれに対応するラベルを示す. 固有表現抽出のためのラベルとして BIO 形式を採用する. BIO 形式とは, 対象となる固有表現のチャンクの最初の要素のラベルの先頭に「B-」,それ以降の要素のラベルの先頭に「I-」, そして, 固有表現として抽出しないチャンクに「O」ラベルを付与する形式である.

本研究では, モデルの学習に使用するため複数のデータセットを用意する. 役者評判記のデジタルテキストから表 1 で示した記号を除去し, 文字コードを統一する前処理を行う. そして形態素解析を行い, 表 2 の抽出項目に対応したラベルのアノテーションを行った単語ベースデータセットを作成する. そしてテキストを文字に分割した文字ベースデータセットを作成する.

表 2 抽出する役者情報と対応するラベル

抽出項目	対応ラベル
役者名	PSN
座本	OWN
役種	ROLE
位付	EVA
評価者	VAL

4.1. 単語ベースデータセット

単語ベースデータセットは, 役者評判記のデジタルテキストに対して形態素解析器にMeCab を使用し，形態素解析を施したものにアノテーションを行なったデータセットである．本研究で使用する役者評判記は 1737 年から 1772 年までの近世に出版されたものである. そのため, 現代語の形態素解析辞書を使用した形態素解析器では高い精度での形態素解析は望めない. そこで形態素解析辞書として小木曽ら[8]によって開発された近世口語（洒落本）UniDic[9]を使用する. 近世口語（洒落本）UniDic は洒落本コーパスを使用し作成されている. 洒落本コーパスに使用されている 5 つの洒落本のうち 4 つが今回使用する役者評判記の発行時期との差が 5 年以内となっている．そして，戸塚[10]の研究結果より，近世口語(洒落本)UniDic は役者評判記に最も有効な形態素解析辞書であることがわかっている．そのため，役者評判記の形態素解析辞書として適していると考える. 形態素解析を施したデータセットに表 2 で示した抽出項目に対応したラベルを付与した.

4.2. 文字ベースデータセット

文字ベースデータセットは, 単語ベースデータセットを機械的に文字に分割し, 各文字に対応したラベルを付与したデータセットである.

5. 提案手法

本研究では，固有表現抽出手法として BiLSTM-CRF モデルを使用する. 実験には, 機械学習ライブラリの Torch[11]で提供されている BiLSTM-CRF モデルを使用した．実験で使用したハイパーパラメータを表 4 に示す．また, 各データセットに対して同じハイパーパラメータを使用している. 本研究で使用する各モデル図を図 5，図 6 に示す.

6. 評価実験

本研究では, 役者評判記の『役者多名卸（江）』, 『役者満友家（江）』を用いて提案モデルの学習を行った. 提案モデルを学習する際に使用するデータを学習用データ 8 割と, 検証用データ 2 割に分割し，学習用データで提案モデルを学習する. 実験の評価指標として再現率（recall）, 適合率（precision），そして再現率と適合率の調和平均である F 値（f-measure）を使用し，それぞれを以下の式 1，式 2，式 3 で求めた.

役者名を表すラベル「B-PSN」を例に式 1, 式 2, 式 3 の変数の説明を行う. TP（True positive）は「B-PSN」であるものを正しく「B-PSN」と予測できた数値である．TN（True Negative）は「B-PSN」ではないラベルを正しく「B-PSN」ではないと予測できた数値である．FP（False Positive）は「B-PSN」ではないものを誤って「B-PSN」と予測した数値である. FN（False Negative）は実際に「B-PSN」であるものを誤って「B-PSN」以外のラベルを予測した数値である.

$$recall = \frac{TP}{FN + TP} \quad (1)$$

$$precision = \frac{TP}{FP + TP} \quad (2)$$

$$F - measure = 2 \times \frac{reacall \times precision}{reacall + precision} \quad (3)$$

単語ベースデータセットを使用したモデルの性能評価結果を表 5 に, 文字ベースデータセットを使用したモデルの性能評価結果を表 6 に示す.

実験の結果, 平均的に単語ベースデータセットを用いたモデルが高い精度を示した. 役者評判記はデジタル化される際に構造化されており, 本研究で定義した抽出する役者情報の単語のみで文を構成しているものが存在する． 具体的には，役種（ROLE），評価（EVA），評価者（VAL）である．このように一定の構造の中で出現する単語に関しての抽出精度は再現率, 適合率どちらかが非常に高い精度を示している．一方, 評価文の中の文脈にも出現する単語である役者名（PSN）や座元（OWN）は，役者情報の単語のみで構成されている文に出現する単語よりも低い精度を示すことが多い結果となった.

7. あとがき

本研究では, 歌舞伎役者の芸評書である役者評判記のテキストに対して固有表現抽出を行う手法を提案し，評価実験を行なった．実験の結果, 役者評判記の固有表現抽出において形態

素解析を用いて文章を分割したデータセットを用いたモデルがより高い抽出精度を示した.

今後の研究では, データセットの量を増やすことを考えている. また, 事前学習で教師なし学習を行う BERT を用いた実験を行うことを検討している.

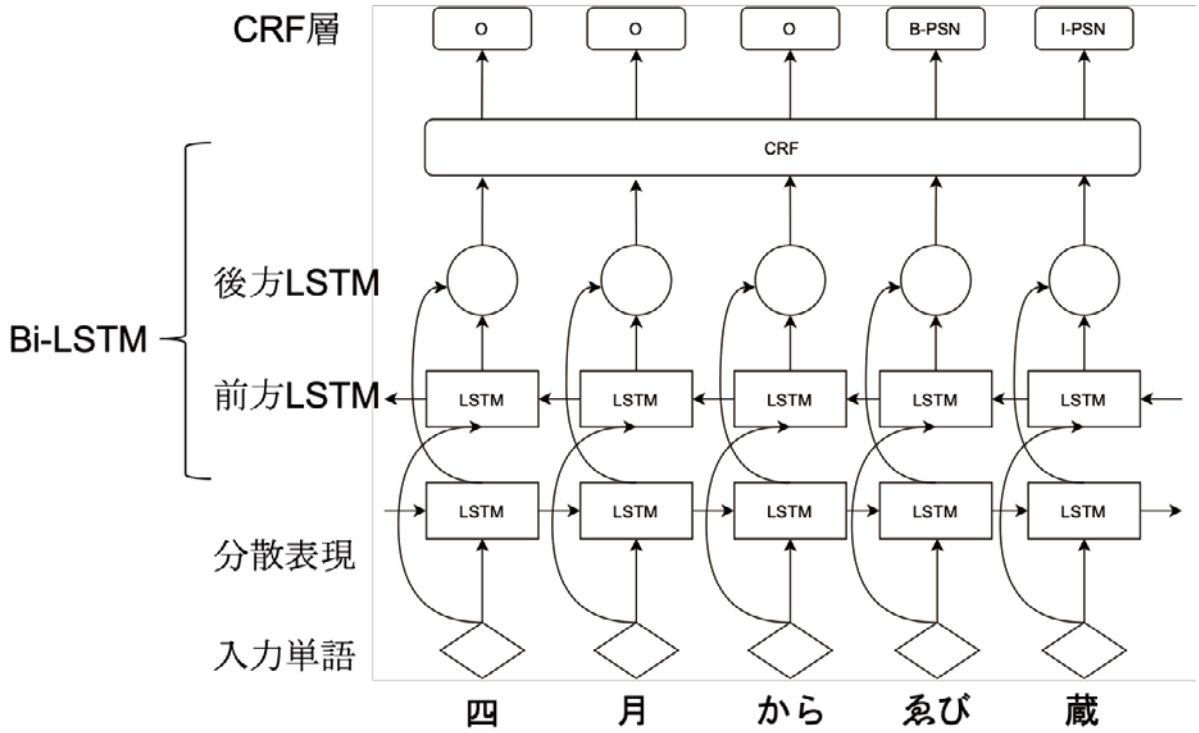


図 5 単語ベースデータセットを使用した BiLSTM-CRF モデル

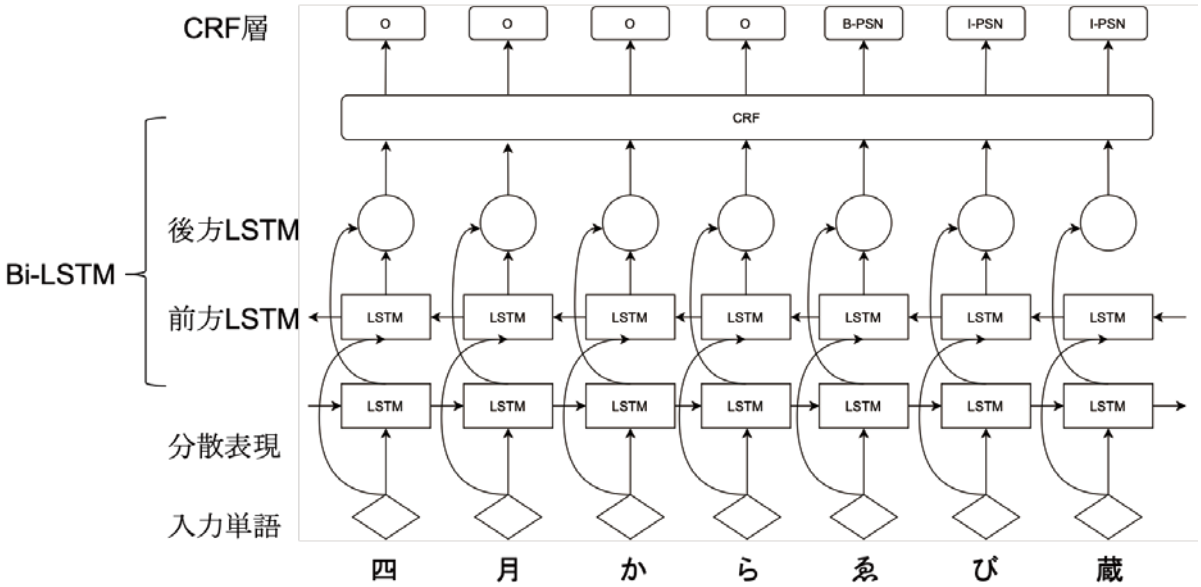


図 6 文字ベースデータセットを使用した BiLSTM-CRF モデル

表 4 使用したハイパーパラメータ

ハイパーパラメータ	値
hidden dimension	10
embedding dimension	50
エポック数	10
学習率	0.01
最適化手法	SGD

表 5 単語ベースデータセットを使用したモデルの評価結果

	再現率	適合率	F値
PSN	80.87	83.04	81.94
OWN	84.85	93.33	88.89
ROLE	90.00	100.00	94.74
EVA	91.18	96.88	93.94
VAL	90.48	95.00	92.68
平均値	87.46	93.65	90.43

表 6 文字ベースデータセットを使用したモデルの評価結果

	再現率	適合率	F値
PSN	77.39	88.12	82.41
OWN	87.88	90.62	89.23
ROLE	80.00	72.73	76.19
EVA	91.18	98.41	94.66
VAL	85.71	100.00	92.31
平均値	84.43	89.97	86.96

謝辞

役者評判記のテキスト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] 山本純子，大澤留次郎：古典籍翻刻の省略化くずし字を含む新方式 OCR 技術の開発，情報管理，2016，Vol.58，No.11，pp.819-827.

[2] 役者評判記研究会， https://www.arc.ritsumei.ac.jp/archive01/theater/document/hyobanki/index-j.htm，（参照 2021-10-29）.

[3] Z. Huang, W. Xu, K. Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991, 2015.

[4] Language-Independent Named Entity Recognition，https://www.clips.uantwerpen.be/conll2003/ner/，（参照 2021-10-31）.

[5] G. Lample，M. Ballesteros，K. Kawakami，and C. Dyer. Neural Architectures for Named Entity Recognition. Proc. of NAACL-HLT, 2016，pp.260-270.

[6] 矢野憲，伊東薫，若宮翔子，荒巻英治，深層学習による医療テキストからの固有表現抽出器の開発とその性能評価，2017，人工知能学会全国大会（第 31 回）.

[7] 白井圭佑，森信介，後藤真：人名辞典からの知識抽出，じんもんこん 2020 論文集, 2020, pp.11-16.

[8] O. Toshinobu, M. Komachi and Y. Matsumoto. Morphological Analysis of Historical Japanese Text, Journal of Natural Language Processing，2013，Vol.20，No.5， pp.727-748.

[9] 近世口語 UniDic，Unidic，https://ccd.ninjal.ac.jp/unidic/download_all#unidic_kinsei，（参照 2021-10-31）.

[10] 戸塚史織：役者評判記計量テキスト分析のためのノート：現状と課題，アート・リサーチ，2021，Vol.21，pp.83-92.

[11] PyTorch，https://pytorch.org/（参照 2021-10-29）.