CRF を用いた児童文学の人物表現抽出

2016SC075　大田那央
指導教員：河野浩之

1　はじめに

公益社団法人全国出版協会の調査 [1] によると，近年，Kindle などの電子書籍などの普及により電子出版の市場は年々拡大していて，2015 年の 1,502 億円から 2019 年までの 5 年間で 3,072 億円と 2 倍に拡大している．このような状況で，電子化された物語に対して情報を抽出して整理することで，長い期間が経って物語を読み返す際に内容の想起を支援するシステムなどが提案されている [4][6]．これらのシステムでは登場人物を抽出して関係性を示したり，登場人物の初登場シーンにジャンプする機能を備えており，これによって，登場人物がどんな人物であったかを把握することができるようになっている．しかし，登場人物を完全に抽出することは困難である．登場人物の表現は作品によって多種多様であるため，未知語や特殊な表現が多く，それらを抽出できるかが課題となっている．

本研究ではこの事に着目し，登場人物の抽出を試みる．児童文学では，人間以外にも動物や果物などが登場人物として現れる場合が多いため児童文学を対象に実験を行う．

2　関連研究

馬場ら [2] は，英米文学の推理小説を対象に形態素解析器を使用し，品詞が「名詞-固有名詞-人名-一般」,「名詞-固有名詞-人名-姓」,「名詞-固有名詞-人名-名」と解析された形態素を人名として抽出している．人名抽出にあたって，辞書を使用して人名抽出の網羅性を高めているが，辞書に存在しない未知の人物や，動物などの登場人物表現に対しては抽出ができない．

米田ら [3] は，“小説内において登場人物は主語として必ず登場する”という仮説をたて，構文解析を用いて，小説中の各文の主語を登場人物の候補として抽出している．しかし，構文解析による文節区切りの結果を利用するため，登場人物を表す表現として「〜のおかあさん」などの表現は「〜の」と「おかあさん」に区切られてしまうため，人物候補とならない．

本研究では，CRF を用いた固有表現抽出器によって高い抽出性能である MA ら [5] の研究を参考にした．CRF とは，入力となるデータ列に対して，個々のデータに全体で最適なタグを付与する機械学習の手法である．MA ら [5] は固有表現抽出器によって，人物と場所表現の抽出を行っている．この研究では，5 つの物語テキストを使用しており，学習と評価の両方に用いて実験を行っていた．本研究では，30 の作品を学習に使用し，学習に使用する作品と評価に使用する作品を分けて学習データに存在しない作品に登場する人物表現がどの程度抽出できるかを検証する．

3　人物表現の抽出手法

人物表現抽出手法の概要について述べる．図 1 に手法の処理の流れを示す．

3.1　タグ付けテキストの作成

機械学習を用いて固有表現抽出器を作成するためには児童文学のテキストを学習に適した形式にする必要がある．学習に使用するテキストをタグ付けテキストとして，作成手順を以下に示す．

1. 児童文学テキストを 1 文ずつ分割して形態素解析を行い，単語と品詞情報を得る．
2. 得た形態素に対して，IOB2 タグ形式で人物表現に対してタグを付与する．

IOB2 では，固有表現を表す始まりの形態素に B タグを付与し，その続きの形態素に I タグを付与し，それ以外の形態素には O タグを付与する．単語の表記，文字種，品詞と品詞細分類，前後 3 つの単語を学習情報として CRF による学習を行い，固有表現抽出器を得る．固有表現抽出器で付与されたタグを元に分割された形態素を繋げることで人物として出力させる．

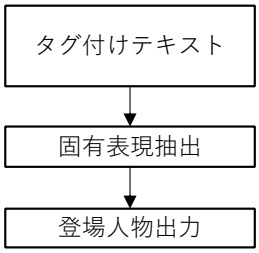

図 1　人物表現抽出手法の概要

これらの手法を行っていくうえで，プログラミング言語には Python3.8.5 を用い，形態素解析器には MeCab を，MeCab の辞書には標準で用意された IPADIC 辞書と固有表現を多く含む mecab-ipadic-neologd 辞書を併用した．CRF を行うライブラリ CRFsuite を用いる．

4　実験

固有表現抽出器の学習に使用した作品群を表 1 に示した．これらの作品は青空文庫 [7] の分野別リストから児童書内の文学作品を収集した．全部で 30 作品であり，分割した文章の合計は 3,535 文である．品詞情報を用いて学習を行う際，誤った品詞での誤学習を防ぐため，登場人物に対する形態素解析に誤りの少ない作品を選択した．

表 2 に示した 5 つの作品は人物表現の抽出に使用した作

表 1 学習に使用した作品

赤いくつ	赤とんぼ	おしゃべり姫
シンデレラ	山姥の話	おやゆび姫
かえるの王様	ブレーメンの町楽隊	ラプンツェル
はだかの王様	ヘンゼルとグレーテル	和尚さんと小僧
一寸法師	ジャックと豆の木	猿小僧
かちかち山	金太郎	浦島太郎
小人のくつやさん	クねずみ	白雪姫
マッチ売りの少女	みにくいアヒルの子	ツぇねずみ
もみの木	ねずみの嫁入り	牛若と弁慶
姨捨山	大男の話	雪女

品である．これらの作品を用いて固有表現抽出器の適合率と再現率の評価および，人物表現の抽出を行った．

表 2 人物表現抽出に使用した作品

手袋を買いに
はなさかじじい
桃太郎
瘤とり
セロ弾きのゴーシュ

表 3 は，表 2 に示した 5 作品の適合率，再現率の平均値と，単純比較することはできないが参考のため先行研究 [5] の抽出結果と合わせて示した．先行研究 [5] では，5 つの作品を学習と評価に使用していたため，本研究で算出したものと比較して再現率が高い．しかし，適合率に関しては，本研究の数値が高くなっている．学習に使用するデータが多くなったため，人物表現が登場する際の周りの単語の情報を多く学習できたためであると考えられる．

表 3 固有表現抽出器の評価結果

	B-CHAR		I-CHAR	
	適合率	再現率	適合率	再現率
5 作品の平均	0.96	0.66	0.98	0.97
MA ら [5]	0.88	0.81	0.98	0.91

表 4 は，手法によって実際に抽出した人物表現である．比較を行うために人手で抽出した人物もあわせて示した．

表 4 抽出した人物表現：手袋を買いに

固有表現抽出器によって抽出した人物表現
’狐の親子’, ’母さん狐’, ’坊やの狐’, ’母さんの狐’, ’狐の子’, ’二匹の狐’, ’トントン’, ’人間のお母さん’, ’帽子屋’, ’狐の子供’, ’子狐’, ’お母さん’, ’子供の狐’, ’お友達の狐’, ’お母さん狐’, ’帽子屋さん’
人手で抽出した人物表現
’狐の子’, ’坊やの狐’, ’母さんの狐’, ’お母さま’, ’お母ちゃん’, ’お友達の狐’, ’お母さん狐’, ’二匹の狐’, ’人間のお母さん’, ’帽子屋’, ’坊や’, ’狐の子供’, ’子供’, ’母さん’, ’子供の狐’, ’狐の親子’, ’子狐’, ’狐’, ’母さん狐’, ’お母さん’, ’母ちゃん’

それぞれ抽出した人物を比較すると，すべての人物表現を正しく抽出することはできなかったが，複数の形態素からなる人物表現や，辞書を拡張して形態素解析器を行っている抽出方法では抽出ができなかった未知語に対する抽出や，動物など人物名ではない登場人物の表現，構文解析による文節区切りでは抽出ができない「人間のお母さん」などの人物表現に助詞が含まれる表現においても，児童文学をある程度学習させた固有表現抽出器であれば抽出が可能であることが確認できた．

「トントン」という単語が誤って抽出されている．「トントン」という単語は，作品中では擬音語であり人物表現ではない．形態素解析では「名詞-固有名詞-人名-一般」と誤解析されていた．擬音語などの形態素解析において誤解析しやすい形態素を除外する処理を加えることで改善する必要がある．

5 おわりに

本研究では，先行研究をもとに CRF という機械学習の手法を用いた固有表現抽出器を作成し，児童文学の人物表現の抽出を行った．先行研究と比較して学習する作品数を増やすことで学習に使用していない作品に対しても人物表現を抽出できること，多様な人物表現が抽出できることが確認できた．

参考文献

[1] 公益社団法人 全国出版協会・出版科学研究所，“2019 年出版市場（紙＋電子）を発表しました，” https://www.ajpea.or.jp/information/20200124/index.html，参照 July 14， 2020.

[2] 馬場こづえ，藤井敦，“小説テキストを対象とした人物情報の抽出と体系化，” 言語処理学会第 13 回年次大会発表論文集，pp．574-577，2007.

[3] 米田崇明，篠崎隆宏，堀内靖雄，黒岩眞吾，”述語情報を利用した小説の登場人物の抽出，” 言語処理学会第 18 回年次大会発表論文集，pp．855-858，2012.

[4] 田中翔太郎，阿部誠，尾内理紀夫，”物語の内容想起支援インターフェースの開発，” 第 55 回プログラミング・シンポジウム予稿集，pp．7-15，July，2014.

[5] MA Jiaxiu，西原陽子，山西良典，“物語内の人物と場所情報の時系列可視化による読書支援，” 情報アクセスと可視化マイニング研究会 (第 23 回)，pp．9-12，2019.

[6] 謝涵，西田健志“物語の登場人物を把握しやすくするシステムの提案，” 情報処理学会研究報告ヒューマンコンピュータインタラクション，pp．1-5，2017.

[7] 青空文庫，https://www.aozora.gr.jp/，参照 July 18，2020.