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

Novel Reading Support by Time Series Visualization of Characters and Positions

MA Jiaxiu[1] *　西原陽子[2]　山西良典[2]
JIAXIU MA[1]　Yoko Nishihara[2]　Ryosuke Yamanishi[2]

[1] 立命館大学情報理工研究科
[1] Graduate School of Information Science and Engineering, Ritsumeikan University
[2] 立命館大学情報理工学部
[2] College of Information Science and Engineering, Ritsumeikan University

Abstract: This paper proposes a time-series visualization method of characters and positions in novel text. The proposed method will support users when they restart reading books. Once the already read part of a book is given, the proposed method extracts characters and positions every sentence and visualize them in time-series. As the intermediate report, we conducted experiments for a method of extraction of characters and positions. The averaged precision was from 0.88 to 0.98 and the averaged recall was from 0.79 to 0.97.

1　はじめに

インターネットの普及にともない，伝統的な紙の本の代わりに電子書籍が流通するようになった．電子書籍は 1 つの端末に複数の書籍を入れて持ち歩くことが可能である．複数の書籍を並行して読むこともしやすくなった．

長い時間を読書に取れない人は，1 つの書籍を複数回に分けて読むこともある．複数の書籍を並行して読んでいると，1 つの書籍の読書を中断してから再開するまでに長い時間がかかることもある．長い時間があいてしまうと，読書を再開するときに読んだ内容を忘れてしまうことがある．忘れたまま読み進めると，書籍の内容を把握しにくくなり，読書を十分に楽しむことができない．特に多くの人物が出てくるような物語であると，どの人物が誰とどこに居たのかが分かりにくくなる．既読部分の人物と場所の情報が可視化され，読書再開前に眺めることができれば，読んだ内容を思い出すことが容易になり，読書の支援ができると考えられる．

そこで本研究では，書籍の中の物語を対象とし，物語の既読部分から人物と場所情報を抽出し，時系列として可視化する手法を提案する．文から人物と場所を表す情報を抽出するだけでなく，人物が存在する場所を特定し，特定された情報を時系列として可視化することに挑戦する．本稿では中間報告として，物語のテキスト内の人物と場所情報の抽出に対する評価結果を報告する．

*連絡先：立命館大学情報理工学研究科
〒 525-0058 滋賀県草津市野路東 1-1-1
nisihara@fc.ritsumei.ac.jp

2　関連研究

小説テキストを対象とした人物情報の抽出手法が提案されている [1, 2]．文献 [1] では，人名辞書「8 万人西洋人名よみ方綴り方辞典」を利用し，辞書に載っている人名を小説テキストから抽出する.「性別」,「年齢」,「年代」,「職業」,「身体的特徴」,「性格」の六種類の人物情報を抽出し，人物リストを作成し，英米文学の推理小説を対象とし，人名の抽出を行う．本研究では人物リストにない人物情報を抽出するために，辞書ではなく，物語のテキストを機械学習することにより抽出を行う．

物語テキスト中の登場人物の関係を抽出する手法が提案されている [3]．人物関係を表す表現の辞書を作成し，表現に合致するパターンを構築し，パターンを用いて抽出を行う．本研究では人物間の関係ではなく，人物と場所の関係を扱う．誰がどこにいたのかの情報を可視化することにより，物語の進行を思い出しやすくする．

物語テキストから進行状況に応じて登場人物の存在状態と関係を推定する手法が提案されている [4]．存在状態を人物の生死に関わる「生存」,「死亡」,「死亡 候補」の三種類とし，判定をする．本研究では登場した

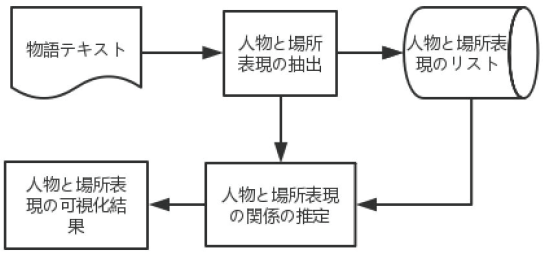

図 1: 提案手法の処理の流れ

人物の存在状態は扱わないが，既存研究と組み合わせることで，より詳細な時系列の可視化が可能になると考えられる．

3 提案手法

提案手法の概要を説明する．図 1 に提案手法の処理の流れを示す．入力は，物語のテキストである．ユーザの既読部分までを入力とする．入力された物語のテキストから，文ごとに人物と場所を表す文字列を抽出する．各文において，人物と人物がいる場所を紐づける．一文ごとに人物と場所の情報を可視化する．

本研究で理想とする可視化結果を図 2 に示す．物語「赤ずきんちゃん」の冒頭から 67 文目までの，人物と場所情報が正確に抽出された場合の可視化イメージ図である．

3.1 人物と場所表現の抽出

物語のテキストから人物と場所を表す表現を抽出する．人物と場所は固有表現の一種である．提案手法では固有表現抽出器の一つである Conditional Random Fields (CRF) [6] を用いる．CRF は条件付確率場のモデルであり，全体で最適な固有表現のためのタグ付けを行う手法である．CRF を用いた固有表現抽出器の作成手順を示す．

1. 物語のテキストを文ごとに分割する．
2. 各文に対して形態素解析を行い，単語と品詞情報を得る．
3. CRF で学習を行うために，単語に対し固有表現抽出のためのタグを付与する
4. CRF を用いて学習を行う．1 つの単語に対し，自分自身と前後 3 つの単語，および品詞情報を学習し，固有表現抽出器を得る．

表 1: 提案手法で用いる BIO2 タグ．人物と場所表現用のタグを用いる．

表現タグ	説明
B-CHAR	人物表現文字列の始まる
I-CHAR	人物表現文字列が続いている
B-POS	場所表現文字列の始まる
I-POS	場所表現文字列が続いている
O	人物と場所表現以外の文字

表 2: 使用した物語のテキスト

番号	タイトル
1	赤ずきんちゃん
2	浦島太郎
3	桃太郎
4	猿蟹合戦
5	良夜

得られた固有表現抽出器を用い，物語のテキストから人物と場所表現を抽出する．物語のテキストは青空文庫などで公開されている物語のテキストを利用する．文末の句読点（．や.），発話終了の鍵括弧など記号があれば文末と判定し，物語テキストの文への分割を行う．形態素解析器は MeCab，辞書は NEologd を用いる．

CRF での学習のためのタグは BIO2 形式 [5] を用いる．BIO2 では固有表現の先頭の形態素に B タグを付け，固有表現内の先頭以外の形態素に I タグを付与する．固有表現以外には O タグを付与する．本研究では人物と場所表現をそれぞれ区別して抽出するため，表 1 に示すタグを用いる．

CRF を用いて学習を行う際は，図 3 に示すように，i 番目の形態素を中心として，前後 3 つの形態素をあわせた合計 7 つの形態素に対し，入力単語，品詞細分類，表現タグを用いる．

4 人物と場所表現抽出の評価実験

人物と場所表現を抽出する部分について適合率と再現率を用いて評価する実験を行なった．

本実験で用いた物語のテキストのタイトルを表 2 に示す．これらのタイトルのうち 1 から 4 は比較的日本人に読まれることが多いものとして選択した．5 のタイトルは，日本人には比較的知らない人が多いと思われるが，比較のために青空文庫の著者名順の先頭のものを選択した．

BIO2 のタグ付けは第一著者が行なった．タグづけされたデータを物語テキストごとに学習し，適合率と再現率を算出した．データは全て学習に用い，学習の際

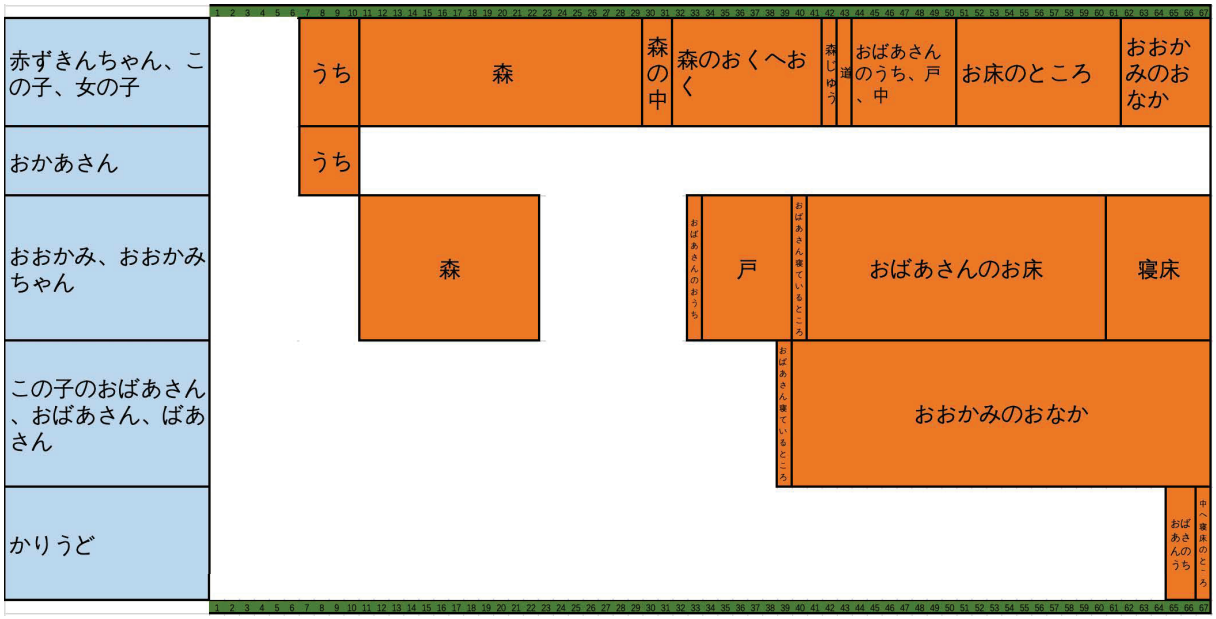

図 2: 物語「赤ずきんちゃん」の提案手法での可視化イメージ図．横軸に文の番号（緑色の長方形），縦軸に人物（水色の長方形），人物の登退場状態（オレンジ色の長方形），内部に人物が居る場所（オレンジ色の長方形中の黒字）を示す．

位置	入力単語	品詞細分類	表現タグ
i-3	これ	名詞-代名詞 -一般	O
i-2	を	助詞-格助詞-一般	O
i-1	赤ずきん	名詞-固有名詞-一般	B-CHAR
i	ちゃん	名詞-接尾-人名	I-CHAR
i+1	、	記号-読点	O
i+2	ここ	名詞-代名詞-一般	O
i+3	に	助詞-格助詞-一般	O

図 3: CRF で学習する際の学習情報の例

の適合率と再現率を算出した．適合率と再現率は $B-CHAR$，$I-CHAR$，$B-POS$，$I-POS$ ごとに算出した．

4.1 実験結果と考察

算出された適合率と再現率を表 3 に示す．適合率の平均は 0.88 から 0.98 になった．再現率の平均は 0.79 から 0.97 になった．

CHAR タグ，POS タグのいずれにおいても，B タグの方が I タグよりも抽出における適合率と再現率が低かった．B タグは固有表現の冒頭の形態素に付与されるが，今回の固有表現抽出では人物表現，場所表現の冒頭の単語や品詞細分類が多様であったために，適合率，再現率がともに低くなったと考えられる．失敗例を表 4 に示す．失敗した例を表 4 に全て書いてもらえますか．物語ごとに分けてこれに対し I タグは B タグが出現した場合に後続するものであり，条件が限定されやすい．このため I タグの適合率と再現率が B タグよりも高くなったと考えられる．

実験に用いた物語のテキストのうち 1 から 4 は適合率と再現率は高かったが，5 のテキストは適合率と再現率が低かった．1 から 4 のテキストは子供向けの物語であった．5 のテキストは「良夜」という日本の物語で，対象とする読者の年齢層は低くない．「良夜」の登場人物は「伯父」，「父」が多く，これらの抽出に失敗することが多かった．人物名が具体的な名称（「赤ずきんちゃん」「浦島太郎」「桃太郎」など）ではないことがあり，人物名と認識がされにくかったと考えられる．

5 おわりに

本研究では，既読部分から人物と場所表現を抽出し時系列で可視化することにより，読書を支援するための手法を提案した．本稿では手法の一部である，人物と場所表現の抽出部分について評価実験を行い，適合率と再現率により評価を行った．実験の結果，適合率の平均は 0.88 から 0.98 になった．再現率の平均は 0.79 から 0.97 になった．子供向けの物語の方が人物，場所表現の抽出が容易であることがわかった．今後は人物と場所の関係を推定すること，人物と場所の時系列可視化を行うことが課題である．

表 3: 人物と場所表現の抽出の適合率と再現率

	B-CHAR		I-CHAR		B-POS		I-POS	
物語	適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
1. 赤ずきんちゃん	0.94	0.90	1.00	0.96	1.00	0.90	0.99	0.97
2. 浦島太郎	0.97	0.98	0.94	0.91	0.98	0.79	0.96	0.97
3. 桃太郎	0.96	0.97	0.95	1.00	1.00	0.89	1.00	1.00
4. 猿蟹合戦	0.98	0.98	1.00	1.00	1.00	0.87	0.97	0.97
5. 良夜	0.55	0.22	1.00	0.68	0.78	0.51	0.78	0.94
平均	0.88	0.81	0.98	0.91	0.95	0.79	0.94	0.97

表 4: 本手法の抽出失敗の結果例（B タグと I タグそれぞれの失敗例を示し，形態素解析の結果は記号”｜”に分けた.）

物語	人物・場所	人手で付けたタグ	手法で付けたタグ
1. 赤ずきんちゃん	おばあさん	B-CHAR	O
	おばあさん｜の｜着物	O｜O｜O	B-CHAR｜O｜O
	森｜じゅう｜かけまわっ	B-POS｜I-POS｜O	B-POS｜I-POS｜I-POS
2. 浦島太郎	浜｜べ	B-POS｜I-POS	O｜O
	海	B-POS	O
	かめの子	B-CHAR	O
3. 桃太郎	川	B-POS	O
	帰り	O	B-CHAR
	陸	B-POS	O
4. 猿蟹合戦	山道	B-POS	O
	かに	B-CHAR	O
	山｜へ	B-POS｜O	B-POS｜I-POS
5. 良夜	父	B-CHAR	O
	新潟｜県｜下	B-POS｜I-POS｜O	B-POS｜I-POS｜I-POS
	伯父	B-CHAR	O
	猿	O	B-CHAR
	母	B-CHAR	O
	東京	B-POS	O

参考文献

[1] 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* 回年次報告, Vol. 13, pp. 574–577 (2007)

[2] 米田 崇明, 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物語テキストを対象とした登場人物の関係抽出, 言語処理学会 第 *18* 回年次大会 発表論文集, Vol. 18, pp. 855–858 (2012)

[3] 西原 弘真, 白井 清昭: 物語テキストを対象とした登場人物の関係抽出, 言語処理学会第 *21* 回年次大会発表論文集, Vol. 21, pp. 628–631(2015)

[4] 縣 啓治, 伊藤 雄一, 高嶋 和毅, 北村 喜文, 岸野 文郎: 物語 テキストから進行状況に応じて登場人物の存在状態と関係を推定する手法, 第 *18* 回インタラクティブシステムとソフトウェアに関するワークショップ ,(2010)

[5] 中野 桂吾, 平井 有三: 日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, Vol. 45, pp. 934–941(2004)

[6] 坪井 祐太, 鹿島 久嗣, 工藤 拓: 言語処理における識別モデルの発展 HMM から CRF まで, 日本 *IBM* 株式会社東京基礎研究所 グーグル株式会社,