物語テキストを対象とした登場人物の関係抽出

西原 弘真　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{nishihara.h.1991, kshirai}@jaist.ac.jp

1 はじめに

多くの人が趣味として読書を楽しんでおり，電車による通勤時間や勤務中の昼休み等を利用して細切れに読書を行うことはよくある．読書を再開する際，物語の状況を瞬時に把握するのが困難な場合には，前の部分を少し読み返す必要がある．このような時でも，読書の再開前にあらかじめ物語の登場人物に関する情報を整理できれば円滑に読書を再開できる．本研究では，物語テキストから登場人物の関係を自動的に抽出する手法を提案する [7]．テキストから物語の登場人物を抽出すると同時に，親子，兄弟などの家族関係，友人，会社の同僚などの仲間関係など，様々な登場人物間の関係も推定する．

2 関連研究

物語文を対象とした研究として，小林は物語をシーンごとに分割する手法を提案した [4]．既存の辞書などを利用して場所，時間，人物候補を抽出し，これら 3 つのカテゴリ別に数えた語句の異なり数を基準としてシーンを分割する．また，米田らは物語から登場人物を抽出する手法を提案した [8]．局所出現頻度と共起する述語情報を利用して未知の人物名の抽出に成功している．

物語文から関係を抽出することに焦点を当てた研究もある．馬場らは人物間の関連度の重みを含む人物関係図を構築した [1]．小説から登場人物を自動抽出した後，人物が発言したか否かの「台詞情報」，人物が特定の場面に存在するか否かの「入退場情報」及び共起頻度を用いて人物間の重みを決定する．神代らは会話文から友好・敵対関係及び上下関係にある人物を抽出した [3]．発話文と話し手の相対的な位置などを素性とした機械学習により話し手と聞き手を同定し，さらに「わたくしめ」のような人称表現などを素性として人物関係の有無を判定する分類器を学習する．Elson らは物語における人物同士がどれほど密接に関わっているかを人物関係図により示した [2]．ルールベース及び発話文の周辺情報を素性とした機械学習により発話者を同定し，発話内にある人物名の言及数や発話外に出現する人物名と発話の距離などを手がかりに人物関係図中のリンクに重みを付与する．Lee らは人物同士及び人物と場所のリンクを含む相関図を構築した [5]．この手法では，人物の関係を表す文を相関図における人物間のリンクに付与する．明示的に人物関係が示されている文はそのまま抽出し，また暗黙的に関係を示す文も FrameNet を用いて抽出する．

馬場らと Elson らの研究では，人物同士がどの程度密接に関わっているかを定量化して示すだけで，人物間の具体的な関係が明示されていないのに対し，本研究では家族関係など具体的な関係の抽出を試みる．神代らの研究では，具体的な関係は示されているものの，関係を 2 種類に限定している．一方，本研究では家族関係や恋人関係など様々な関係の獲得を試みる．Elson ら，神代らの研究は発話文に着目しているのに対し，本研究では地の文から関係の抽出を試みる．また，Lee らの研究は本研究と内容が近いものの，人物間の関係がテキストから抽出された文のままで提示されるので，一目で人物関係が把握しづらい点が問題となる．これに対し，本研究では「太郎 - 姉 - 花子」のように関係を定式化して提示する．また日本語で書かれた物語を解析の対象とする．

3 提案手法

図 1 に提案手法における処理の流れを示す．まず，人物関係を表す語を収録した関係辞書と，文から人物間の関係を抽出する関係抽出パターンの集合を事前に構築する．小説テキストに対し形態素解析などの前処理を行い，登場人物を抽出し，登場人物リストを作成する．登場人物リストと関係辞書を参照し，人物間の関係を表す可能性のある文 (以下，人物関係文と呼ぶ) を抽出する．最後に，人物関係文から関係抽出パターンを用いて人物関係を抽出する．抽出する人物関係は以下のいずれかとする．

3 項関係　P1 - R - P2　(ex. 太郎 - 妹 - 花子)　(1)

2 項関係　P1 - P2&R　(ex. 太郎 - 父)　(2)

P1,P2 は登場人物，R は関係辞書における関係語を表す．2 項関係における P2&R は人物と関係の両方を表す語である．例えば，「太郎の父」という句における「父」は，ある人物の存在を示すと同時に，その人物は太郎の「父」であるという関係が成立することを表す．本研究ではこのような 2 項関係も抽出の対象とする．

本論文で論ずるのは人物関係の抽出までであるが，将来的には抽出した人物関係から小説の人物相関図を構築することも視野に入れている．

3.1 前処理

前処理では，まず物語テキストを Mecab，Cabocha を用いて形態素解析，文節の係り受け解析を行う．

次に，以下の 3 通りの方法で登場人物を抽出する．1 つ目は，Cabocha による固有表現解析で PERSON と認識された語を人物とする．2 つ目は，日本語語彙大系

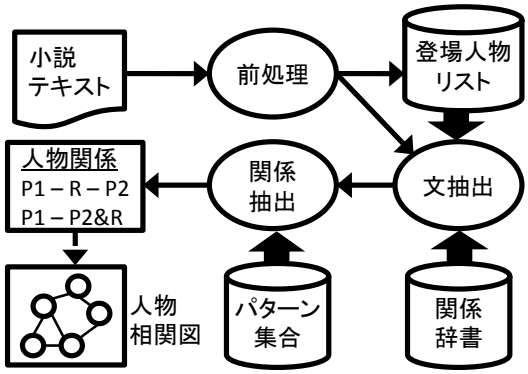

図 1: 提案手法の概要

における「人名」「人」のカテゴリに含まれる語を人物として抽出する．3 つ目は，日本語語彙大系の格フレーム情報における格要素の選択制約が「人」である語を人物とする．例えば，入力文が「ネイマールがペレに憧れる。」のとき，動詞「憧れる」の格フレームが「N1 が N2 に憧れる」であり，N1 の選択制約が「人」の場合,「ネイマール」のような日本語語彙大系には載っていないような人物も抽出できる．これらの処理において，人物ストップワードに該当する語は人物として抽出しない．人物ストップワードは，500 編の小説を対象に登場人物を抽出し，出現頻度 10 回以上の語のうち人物に該当しないものを人手で選択して作成する．

また，より多くの人物関係を抽出するため，省略語を補完するゼロ照応解析を行う．本研究では Nariyama の手法 [6] を実装した．

3.2 関係辞書の構築

関係辞書とは,「父」「同僚」など人物間の関係を示す語を収録した辞書である．本研究では，角川類語新辞典と日本語語彙大系から，人物関係を表す語を人手で選別して構築した．関係辞書は階層構造を持ち，その最上位のカテゴリは「愛憎」「親族」「仲間」「地位」の 4 つである．また，関係辞書では曖昧な関係語とそうでない語を区別する．曖昧な関係語とは，常に関係を表すとは限らない語である．例えば,「主人」は夫という関係を表す場合もあれば単に家の主を指す場合もあるので，曖昧な関係語である．以下，曖昧な関係語を含まない辞書を関係辞書 A，曖昧な関係語を含む辞書を関係辞書 B と呼ぶ．関係辞書 A,B の登録語数はそれぞれ 1287,1735 である．

3.3 人物関係文の抽出

以下の条件を満たす文を人物関係文，すなわち人物関係を表す可能性のある文として抽出する．

1. 登場人物リストにある語を 2 つ以上含み，かつ関係辞書中の関係語を 1 つ含む．
2. 登場人物リストにある語を 2 つ以上含み，そのうちの一つは関係辞書中の関係語でもある．

表 1: 人手構築した抽出パターン

タイプ	パターン
説明型 (節内)	P1 の P2&R [太郎の姉]
	P1 の R (の | は) P2 [太郎の姉の花子]
	P2 (が | は | も) 〜P1 の〜R [太郎は花子の昔ながらの友達]
	P1 (が | は | も) 〜R の P2 [太郎は執拗に妹の花子に]
	P1 (に | には) P2 という R [太郎には花子という娘がいる]
	P1 が〜R、P2 [太郎が友人、次郎を]
行為型 (文内)	P1 (が | は | も) 〜P2&R (と | を | に | の) [太郎が公園で妹と]
	P2&R (が | は | も) 〜P1 (と | を | に | の) [姉が花子を]

※ [ ] 内は各パターンにマッチする文の例

例えば,「葉巻太郎、次郎の兄弟が雨戸をあけると、立っているのは血まみれの暁葉子である。」という文は,「葉巻太郎」「次郎」という 2 名の人物と「兄弟」という関係語が存在するので人物関係文となる．

3.4 パターン集合の構築

人物関係文からパターンマッチにより人物間の関係を抽出する．ここでは，関係抽出のパターンを人手で構築する手法と半自動獲得する手法について述べる．

3.4.1 人手によるパターン構築

10 編の小説から取得した人物関係文を調べ，関係抽出パターンを人手により構築した．パターンは，人物間の関係を直接的に説明している説明型と，2 人物が同じ動詞の項になることで間接的に人物間の関係を示す行為型に分類される．説明型は節内でのマッチングを行い，行為型は文内でのマッチングを行う．表 1 に構築したパターンを示す.「〜」は任意の文字列を表し,「|」は複数の助詞のいずれかにマッチすることを表す．P1, P2, R にマッチした登場人物ならびに関係語を抽出し，式 (1),(2) の人物関係を得る．

3.4.2 パターンの半自動獲得

パターンのテンプレートを用意し，訓練データの小説の集合からテンプレートにマッチするパターンを自動獲得する．自動獲得されたパターンのうち信頼度の高いものを採用する．ただし，信頼度の算出は人手で行うため，半自動獲得手法と呼ぶ．

抽出パターンのテンプレートの一部を表 2 に示す．左辺はパターンマッチの条件，右辺は抽出する 2 項関係もしくは 3 項関係を表す．i:X, $case$ (i=1,2,3) は文節にマッチするパターンを表し，X は文節内に人物 (P) または関係語 (R) が出現することを，$case$ は文節内に出現する助詞を表す．一方，[i,j,syn] は文節 i と文節 j に syn という係り受け関係が成立するという条件を表す．

表 2: パターンのテンプレート (一部)

1:P,*case* 2:*R*,*case* [1,2,*syn*] ⇒ P-*R*
1:P1,*case* 2:P2,*case* 3:*R*,*case* [1,2,*syn*] [2,3,*syn*] ⇒ P1-*R*-P2
1:P2,*case* 2:*R*,*case* 3:P1,*case* [1,3,*syn*] [2,3,*syn*] ⇒ P1-*R*-P2

表 3: 係り受けの条件

syn	定義
direct	文節 i が文節 j に直接係る
share	文節 i と文節 j の直接の係り先が同じ
indirect	文節 i が文節 j に間接的に係る [1]
pred	文節 i と文節 j が間接的に同じ用言に係る

R, *case*, *syn* はテンプレートの変数である．*R* は関係辞書に登録された関係語，*case* は助詞 (助詞が存在しない文節のときは ϕ)，*syn* は表 3 に示した 4 つのいずれかの関係が埋められる．実際には，表 2 のテンプレートだけでなく，P1，P2，R(2 項関係の場合は P と R) の順序を入れ換えたテンプレートも用意する．例えば,「葉巻太郎、次郎の兄弟が...」という文にテンプレートを適用すると以下のパターンが獲得される．このパターンでは，P1 と P2 は人物にマッチする変数である．

1:P1, ϕ 2:P2,ノ 3:兄弟,ガ [1,2,direct] [2,3,direct]

⇒ P1 - 兄弟 - P2

さらに，得られたパターンを以下の 2 つの手法により一般化したパターンも獲得する．(1) *case* をワイルドカード * に置き換える．(2) *R* を R(c) に置き換える．c は関係語 *R* の関係辞書における上位カテゴリ (愛憎，親族，仲間，地位のいずれか) であり，R(c) はカテゴリ c に属する任意の関係語を表す．

最後に，上記の手続きで得られたパターンにおける候補の信頼度を求める．訓練データの小説集合に対してパターンにマッチする文の数を調べ，2 項関係では 20，3 項関係では 3 個未満の文にしかマッチしないパターンを除外する．次に，パターンにマッチした文において人物関係が成立するかを人手でチェックし，人物関係が成立する文の割合をそのパターンの信頼度とする．ただし，2 項関係のパターンは 20 個の文をランダムに選択して信頼度を算出する．信頼度が閾値 T 以上のものを最終的な抽出パターンとする．関係辞書として辞書 B を用い，500 編の小説を訓練データとしてパターンを半自動獲得したところ，$T = 0.6$ のときに 111 個のパターンを得た．

4 評価実験

評価用データとして，青空文庫 [2] における 10 編の小説を用いた．評価用データは，人手でパターンを構築したときに参照した小説やパターンを半自動獲得する際に訓練データとして使用した小説と異なるものを選択した．評価用データから人手で人物関係を抽出し，正解データとした．人物関係を抽出する際には，文に直接書かれている関係だけではなく，読者が推測できる関係も正解として抽出した．また，複数の文から同一の人物関係が抽出されるときは，まとめて 1 つの関係とみなした．1 つの小説当たりの正解関係数は平均で 25.2 であった．

[1]文節の係り受け関係を 2 回以上辿って到達することを表す．

[2]http://www.aozora.gr.jp/

表 4: 実験結果

	手法	精度		再現率		F 値	
		A	B	A	B	A	B
完全一致	$M_{\text{人手}}$	.270	.223	.226	.258	.246	.239
	$M_{\text{半自動}}$	.536	.495	.179	.198	.268	.283
	M_{B}	.051	.040	.373	.433	.089	.073
部分一致	$M_{\text{人手}}$	.332	.271	.278	.313	.302	.290
	$M_{\text{半自動}}$	.631	.594	.210	.238	.315	.340
	M_{B}	.056	.044	.413	.480	.099	.081

提案手法の性能を人物関係抽出の精度，再現率，F 値 (10 編の小説のマイクロ平均) で評価した．また,「完全一致」と「部分一致」の 2 つの評価基準を用いた.「部分一致」では，3 項関係が正解のとき，その一部の 2 項関係が抽出できたときでも正解とみなす．例えば，正解が「清二 - 嫂 - 高子」のとき,「清二 - 嫂」という関係が抽出されれば正解とする．

人物関係文から無条件に人物関係を抽出する手法，すなわち人物 2 名と関係語が同一文にあれば無条件に関係を抽出する手法をベースラインとし，提案手法と比較する．以下，人手で構築したパターンを用いる手法を $M_{\text{人手}}$, 半自動獲得したパターンを用いる手法を $M_{\text{半自動}}$, ベースラインを M_{B} と記す．

4.1 実験結果と考察

前処理としてゼロ照応解析を行ったが，解析の誤りが多く，ゼロ照応解析をしたときとしないときの比較では後者の方がやや性能が高かった．そのため本論文ではゼロ照応解析をしないときの実験結果のみを報告する．実験結果を表 4 に示す．表中の A,B は関係辞書 A，B を使用したことを表す．

ベースラインは再現率が高いが精度は著しく悪い．それに比べて提案手法 $M_{\text{人手}}$，$M_{\text{半自動}}$ は精度と再現率のバランスが取れており，F 値でもベースラインを上回った．$M_{\text{人手}}$ と $M_{\text{半自動}}$ を比較すると，両者の F 値は 3 割程度で，$M_{\text{半自動}}$ の方が若干高かった．精度は $M_{\text{半自動}}$ の方が良く，再現率は $M_{\text{人手}}$ の方が良い．2 つの関係辞書を比較すると，$M_{\text{半自動}}$ では曖昧な関係語を含む辞書 B を使った方が F 値が高いのに対し，$M_{\text{人手}}$ では辞書 A の方が高い．これは，半自動でパターンを獲得する場合には関係語毎にパターンを獲得するため，パターンの信頼度に関係語の信頼性も反映されているためと考えられる．例えば，辞書 B のみに含まれる「女」は，恋人関係よりも一般の女性を指す場合が多く，$M_{\text{半自動}}$ では「女」を含

むパターンの信頼度が低くなって除外されるのに対し，$M_{\text{人手}}$ では「女」を含む関係が誤って抽出される．

$M_{\text{半自動}}$ において，パターンの信頼度の閾値 T を 0.5 から 1 まで 0.1 間隔で変化させたときの F 値を調べたところ，完全一致では 0.137 から 0.283 まで，部分一致では 0.181 から 0.340 まで変動し，$T = 0.6$ のときの F 値が最大となった．表 4 の $M_{\text{半自動}}$ は $T = 0.6$ のときの結果である．閾値 T は本来なら開発データを用いて最適化するべきである．とはいえ，T をどのように設定しても F 値はベースラインよりも高いことを確認した．

4.2 エラー分析

4.2.1 誤抽出の分析

誤抽出の主な要因を以下に示す．[ ] 内は左から順に $M_{\text{人手}}$，$M_{\text{半自動}}$ における各要因の占める割合である．

1. 人物抽出の誤り [0.21][0.27]
2. マッチしたパターンが不適切 [0.50][0.24]
3. 不適切な関係語の抽出[3] [0.07][0.15]
4. 人物が特定できない関係 [0.11][0.10]
5. 3 項で抽出されるべき関係 [0.09][0.19]
6. 否定された関係や将来的な関係など，実際は成立していない関係の誤抽出 [0.01][0.05]
7. 「師匠-弟子」のような自明な関係の誤抽出 [0.01][0]

両手法とも 1. のエラーが多く，人物抽出の精度を上げる必要がある．$M_{\text{人手}}$ では 2. が一番多かった．人手で作成した 8 種類のパターンは条件が緩いために誤った関係が抽出された場合が多かったと考えられる．3. は曖昧な関係語が物語において関係を表すかを判定する手法を導入する必要がある．4. は，例えば「…昔私が通っていた小学校や、その学校の前から街道続きで、昔の藩主の城跡や、仲間とよく遊んだ老松の海風に…」からシステムは「私 - 仲間」を抽出するものの，「仲間」が具体的に誰を指すかは分からないため正解としなかった．出現回数が少ない登場人物は除外するなど，特定の人を指さない人物表現を抽出しないような工夫が必要となる．5. は完全一致の判定で正解としなかった関係である．6. は文内の否定表現や時制表現を考慮する必要がある．7. は自明な関係のリストをあらかじめ構築しておくことで回避できる．

4.2.2 抽出漏れの分析

抽出漏れの主な要因を以下に示す．

1. 人物が抽出できていない [0.33][0.31]
2. マッチするパターンがない [0.51][0.56]
3. 関係語が辞書に存在しない [0.07][0.04]
4. 関係が暗黙的 [0.09][0.08]

[3]主に辞書 B を使用するために生じる．

両手法とも人物の抽出漏れが目立っていた．人物抽出における再現率の低さの他に，$M_{\text{半自動}}$ においては 1 つの文節から 1 人の人物しか抽出しないため，1 文節に 2 名以上人物がいると関係が抽出されないことも原因のひとつである．2. は抽出パターンの不足が原因であるが，$M_{\text{半自動}}$ においては，信頼度の閾値を下げたり係り受けの条件を無視するように条件を緩めれば抽出できた事例もあった．3. に関しては辞書における関係語を増やす必要がある．4. については現段階では対応が難しい課題である．例えば，「甲州屋の息子と倉田屋の姉娘とのあいだには、半七が睨んだ通りの関係が結びつけられていた。」という文から「お紋 - 恋人 - 藤太郎」という関係を抽出する場合を考える．小説全体を読むと，この文における「睨んだ通りの関係」が恋人関係ということが分かるものの，これを自動的に判定するには高度な言語処理や推論が必要となる．また，「甲州屋の息子」が「藤太郎」を，「倉田屋の姉娘」が「お紋」を指すという照応解析を必要とすることも関係抽出を困難にする要因といえる．

5 おわりに

本研究では物語文から自動的に登場人物の関係を抽出するための一手法を提案した．人物関係抽出の F 値は 0.340 となり，ベースラインを上回った．また主な誤りの要因を調査し，その対応策について論じた．これらの対応策を実現する方法を探究することが今後の課題となる．また，パターンの信頼度を自動的に推定し，パターンを完全に自動構築する手法の確立も重要な課題である．

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 回年次大会発表論文集, Vol. 13, pp. 574–577, 2007.

[2] David K. Elson, Nicholas Dames, and Kathleen R. McKeown. Extracting social networks from literary fiction. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 138–147, 2010.

[3] 神代大輔, 高村大也, 奥村学. 物語テキストにおけるキャラクタ関係図自動構築. 言語処理学会第 14 回年次大会発表論文集, Vol. 14, pp. 380–383, 2008.

[4] 小林聡. 場・時・人に着目した物語のシーン分割手法. 情報処理学会自然言語処理研究会, pp. 25–30, 2007.

[5] John Lee and Chak Yan Yeung. Extracting networks of people and places from literary texts. *In Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pp. 209–218, 2012.

[6] Shigeko Nariyama. Grammar for ellipsis resolution in Japanese. *In Proceedings of the 9th International Conference on Theoretical and Methodological issues in Machine Translation*, pp. 135–145, 2002.

[7] 西原弘真. 物語テキストを対象とした登場人物の関係抽出. Master's thesis, 北陸先端科学技術大学院大学, 3 2015.

[8] 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, Vol. 18, pp. 855–858, 2012.