物語における登場人物の親しさ推定

小井出慎 古宮嘉那子 佐々木稔 新納浩幸
茨城大学工学部情報工学科
{ 11t4076r, kanako.komiya.nlp}@vc.ibaraki.ac.jp
{minoru.sasaki.01, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1 はじめに

本研究では物語の登場人物間の親しさの推定を行う．まず，電子化された物語テキストから登場人物と会話を抽出し，同時に抽出した情報により誰と誰が会話をしているかを推定する．登場人物の抽出には，人手による労力を減らすため，機械学習を用いない手法を利用した．抽出した会話内の敬語の使用程度，また，登場人物間の会話量により，登場人物の親しさを推定する．

2 関連研究

物語テキストに関する既存研究には以下のようなものがある．まず，西原ら[1]の研究では登場人物の関係を自動的に抽出する手法について述べられている．その際，登場人物も自動的に抽出した上で関係抽出を行っており，人物抽出の手法についても紹介されている．他にも，米田ら[2]の研究では述語情報と局所出現頻度を利用して登場人物を抽出している．

教師ありで登場人物間の会話を推定する手法として，神代ら[3]の手法がある．これは登場人物の関係図を自動構築する研究であるが，その中で機械学習を用いた会話の話し手・聞き手を同定する手法について述べられている．また，登場人物間のソーシャルネットワークを抽出する手法として、Elson ら[4]のものがある．これは，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示す研究であるが，会話数が増えるとその人物間は密接であると述べられている．

本研究では，機械学習を用いない発話者同定を行う．更に Elson らの，人物間の会話量が多ければ密接な関係を持っているという事に着目し，登場人物間の親しさを推定する．また，敬語の使用量を利用した登場人物間の親しさ推定も行った．

3 手法

図 1 に入出力における処理の流れを示す．まず物語テキストの整形を行い，解析しやすい形にする．整形した物語テキストに対し形態素解析を行い，登場人物と会話文を抽出する．この際，形態素解析した結果から物語テキストを再構築し，抽出した登場人物と合致する部分にタグ付けを行う．再構築したテキストより発話者を推定，登場人物間の親しさについて出力する．

親しさについては定義することが困難であるため，本研究では相対的な評価で推定を行う．

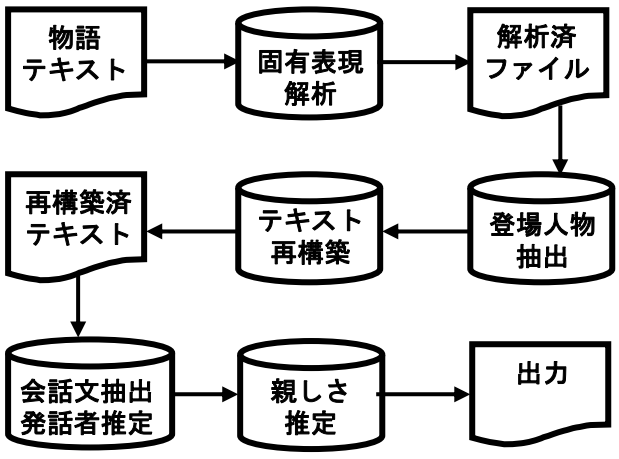

図 1．入出力処理の流れ

以下に，それぞれの処理について詳細を示す．

3.1　登場人物抽出

KNP で固有表現解析を行うことで登場人物抽出を行った．読み込んだ物語テキストを解析し，PERSON タグの付与された固有表現を登場人物として抽出する．

3.2　物語テキストの再構成

抽出した登場人物と解析結果から，物語テキストを登場人物に PERSON タグを付与した上で一文ごとに再構築する．こうすることで，発話者を推定しやすくしている．

3.3　会話文の抽出と発話者推定

カギ括弧(「」)で囲まれた部分を会話として抽出する．会話が 3 文以上発生していない場合，場面の切り替わりと判定する．表 1 に発話者の推定パターンを示す．ここで，P は人物名を示している．また，推定するにあたり話者に優先度を設けた．これは，信頼出来ると思われる話者に高い優先度を設ける事により，精度の向上を図るためである．

まず，発話本文内に着目し

- （一人称）+（は|が）+（人物名）

のように，自らを名乗っていると思われる部分を探索する．稀に（人物名？）+の+（人物名）のように人物名と検出される地名の後に名乗る場合があり，そこは例外検出とした．自らを名乗る場合のパターンは比較的少なく，信頼できると思われるため，最も高い優先度を設定した．

次に，神代ら[3]の論文中にある話し手・聞き手同定実験のベースラインを参考にし，発話の型を以下の 2 つに分け人物名を探索する．

- 組み込み型：発話 1 文後→発話 1 文前→発話 2 文前→発話 2 文後→発話 3 文後
- 独立型：発話 1 文後→発話 1 文前→発話 2 文後→発話 2 文前→発話 3 文後→発話 3 文前

ここで，『「○○○」と A は言った。』『A は頷くと「○○○」と答えた。』のように発話 1 文後の文頭に”と”がある場合のみ組み込み型と判定し，その他を独立型と判定している．また，1 文後が発話であり（人物名）+（敬称）が文末にある場合に限り，聞き手が話し手に呼びかけている可能性が高いと思われるため，検出対象としている．その他の場合は発話内に話し手の名前が出現する可能性は低いと考え，発話は探索対象外としている．発話本文から探索している文の距離が離れるごとに，得られた人物名の信頼は低いと思われるため，優先度は低下させていく．

表 1 発話者推定パターン

	推定パターン	優先度
発話本文	（一人称）（は|が）P 「私が山田太郎です。」	10
	（一人称）（は|が）P(？)の P 「私が山口の山田太郎です。」	9
発話 1 文後	P（敬称）（句読点・感嘆符)」 「山田さん。」	7
	組み込み型	8
発話 n 文	組み込み型	6－n
	独立型	4－n/2
該当無	Noname	0

以上の探索方法で見つからなかった場合，2 文前の発話の人物名を優先度ごと挿入する．これは，会話は一対一で交互に行われているという Elson ら[4]の理論に基づいている．該当しなかった場合は優先度最低の Noname を挿入する．

最後に発話者推定を行った会話文について優先度の比較を行う．前後 2 文の発話者を比較し，優先度が高ければその発話者を挿入する．その際，元々推定した発話文との距離が離れる程信頼は低くなるため，優先度を減少させている．

3.4　親しさ推定

抽出した会話文と推定した発話者から，登場人物間の親しさを推定する．評価方法は予め作品につい

てアンケートを取り,その結果から正解を作成する.また，親しさについては登場人物間の会話量と敬語の使用比率から測る．まず，発話単位で形態素解析を行い，そこから敬語の使用比率を測定する．対象とする敬語は古宮ら[5]の研究を参考に選出した．記号「。」「！」「？」と接続助詞，通常動詞の使用量に対する敬語（ですます，尊敬動詞，謙譲動詞）の比率を求める．この時，発話内に伝聞（『』で始まる別の人物の発言)が存在する場合がある.この場合は，その発話者から聞き手に対する親しさとは異なると考え，対象から除外した．また，その際，発話発生から直近の自分以外の発話者を聞き手とした．この敬語の使用率の算出は，3.3 節で述べた場面の切り替わりごとに行う．これは，同じ章内でも場面ごとに登場する人物は変動するためである．これらの結果を物語の章ごとにまとめ，登場人物間の親しさを算出する.

4　実験と結果

実験の入力データには，青空文庫にて公開されている小説 5 編を用いた.表 2 に使用した小説を示す.小説は解析しやすくするように,予め整形を行った.

4.1　発話者推定実験

3.3 節で述べた条件で会話文を抽出した後，発話者の推定を行った．表 3 に各小説の発話者正答率を示す．正答率は人手で判別を行い，正解した発話/全体の発話で計算した.

4.2　親しさ推定実験

親しさ推定では相対評価を行うため、計 4 名に実験に表 2 の小説を読んでもらいアンケートを集

表 2 実験に使用した小説

まだらのひも	赤毛連盟
白銀の失踪	秘密の庭
金の十字架の呪い	

表 3 各小説の発話者正答率

作品名	正答率[%]
まだらのひも	59.56%
赤毛連盟	55.56%
白銀の失踪	44.34%
秘密の庭	44.85%
金の十字架の呪い	47.27%

表 4 アンケートで使用した項目

親しさ	関係性
1（初対面） ～ 5（よく知っている）	兄姉，弟妹，親，子，教師，弟子，利害，親友，友人，知人，店員，同僚，上司，先輩，後輩

計した．アンケート項目は，A→B への親しさを初対面からよく知っているかの五段階評価と，該当する関係性にチェックする形とした．表 4 に使用した関係の一覧を示す．これらは，物語進行上の章末時点ごとに評価を行った．また，A→B だけではなく B→A の双方を評価する.

今回の実験では，アンケートによって各作品に対する親しさの正解を作成した．この正解は，アンケート内の親しさの項目のみを使用している.そして,最も発話者推定精度の高かった「まだらのひも」を基準とし，敬語の使用率と発話数が最も近かった人物の親しさをそれぞれ付与した．他 4 作品の敬語の使用率と発話数について，アンケートによって作成した正解との平均二乗誤差を求めて結果を比較した.表 5 に計算した結果を示す.

表 5 平均二乗誤差による手法の比較

敬語の使用率	発話数
4.990	5.731

5　考察

発話者推定実験は，最高正答率が 59.65%，最低

が 44.34%という結果となった．これは，発話の前後に登場人物名が出現している作品はそれなりの精度で推定可能であるが，”彼”や”私”の様に代名詞が頻出する場合には対応していないためである．また，KNP の固有表現解析のみを用いているため，人物名の誤検出や検出漏れが発生してしまった．しかし，西原ら[1]の研究のように別の辞書も用いて人物検出の精度を向上させることや，発話者推定時の優先度について，更に他の小説を参照することにより厳密な調整を行う等の方法で更なる精度向上に繋がると思われる．教師なしの手法でも十分な精度が得られれば労力の削減となり，有用な手段となりえるだろう．

親しさ推定実験では，敬語の使用率と発話数から親しさを推定する場合，敬語の使用率から推定する手法の方が良いということが判明した．しかし，敬語の使用率と発話数が最も近かった人物の親しさを付与したところ，敬語の使用率が減るごとに親しさは単調増加するものと想定していたが，実際は離散的になってしまった．これは発話数から推定する手法も同様であるが，事前に行った発話者推定の精度が高くないためであると考えられる．また，登場人物の立場や年齢，性格からも敬語の使用率が低くなることが考えられ，敬語の使用率が低ければ親しいとは言い切れない事も考えられる．敬語使用率の他に，更に別の要素を推定に組み込めれば，より正確な親しさを推定する事が可能であると思われる．

6　おわりに

本研究では，物語テキストに登場する人物間の親しさについて推定した．登場人物に関しては KNP を用いた教師なしの手法で検出，推定を行った．

親しさ推定実験では，アンケートを利用した親しさの正解を作成し，その結果と推定結果との平均二乗誤差を求めることで評価を行った．その結果敬語の使用率を利用した手法の方が，発話数を使った場合よりも精度が良かった．しかし，前段階の発話者推定が 50%前後の精度にとどまっているため，登場人物によっては敬語の使用率が直接親しさには繋がらない場合があったと思われる．そのため，今後の課題としては，より高い精度での発話者推定システムの構築が必要であると考えられる．

謝辞

文部科学省科学研究費補助金［若手 B（No：15K16046）］の助成により行われた．ここに，謹んで御礼申し上げる．

参考文献

[1] 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 pp．380-383，2015.

[2] 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.

[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 pp．628-631，2008.

[4] 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.

[5] 古宮嘉那子，小林明子，乾伸雄，小谷善行，決定木学習による敬語の選択ルールの生成．情報処理学会第 67 回全国大会（平成 17 年）講演論文集第二分冊 1ZA-3，pp．429-430（2005,03,02）