小説テキストに対する登場人物アノテーション

大島一海 [1]　小川浩平 [1]　佐藤理史 [1]
[1] 名古屋大学大学院工学研究科
oshima.kazumi.u4@s.mail.nagoya-u.ac.jp　k-ogawa@nuee.nagoya-u.ac.jp
sato.satoshi.g9@f.mail.nagoya-u.ac.jp

概要

小説理解において，登場人物の認識とセリフの話者推定は，最初の重要なステップである．本論文では，登場人物認識とセリフの話者推定を統合した処理を，**登場人物アノテーション**と名づけ，これを実現するシステムを提示する．さらに，そのシステムを用いて完全自動の登場人物アノテーションを実行し，セリフの話者推定において，主要登場人物リストを与える先行研究と同程度の精度を達成した．

1　はじめに

物語を理解することは，人間にとってはそれほど難しくない．しかし，その描写や構造は，因果推論や常識知識などの様々な要素が入り組み，複雑であるため，物語の理解には高度な情報処理能力が求められる．これを機械的に実現することは，人工知能研究における挑戦的な課題のひとつである．

物語理解には，登場人物の認識，セリフの話者推定，登場人物間の関係抽出などいくつかの段階や種類の処理が想定される．これらのうち，登場人物の認識とセリフの話者推定は，物語理解の入口となる重要な処理である．

我々が提案した登場人物認識システム [1] では，小説テキストに出現する，登場人物を指し示す表現 (メンション) に人物 ID を持つタグを付与する．この認識結果を入力として，テキスト中のセリフにその話者の人物 ID を付与すれば，タグ付与という統一的な形式で，登場人物認識とセリフ話者推定を統合できる．我々は，この 2 つを統合した処理を，**登場人物アノテーション**と名づける．

本論文は，まず，登場人物アノテーションの具体的内容を説明したのち，それを実現するシステムを提示する．さらに，そのシステムを用いて，登場人物アノテーションがどの程度自動化できるかを実験的に明らかにする．

陽の落ちかけた教室で、旧友の<m cid=1>福部里志</m>にそんな意味のことを話した。すると<m cid=1>里志</m>はいつも浮かべている微笑みをちっとも崩さず、言ったものだ。
<u cid=1>「そう思うよ。ところで<m cid=1>僕</m>は<m cid=2>ホータロー</m>に自虐趣味があるとは知らなかったね」</u>
いかにも心外だ。<m cid=2>俺</m>は抗議した。

図 1　登場人物アノテーションの例 (原文: 文献 [2], p. 7)

2　登場人物アノテーション

登場人物アノテーションとは，小説テキストに，登場人物の情報を明示的に付与する作業を意味する．それぞれの登場人物に固有の ID を割り当て，テキスト中の以下の要素に，タグ形式で ID を付与する．

1. 登場人物を指し示す表現 (メンション)
   メンション文字列を m タグで囲い，その cid 属性に，その人物の ID を付与する．
2. セリフ
   セリフ全体を u タグで囲い，その cid 属性に，そのセリフの話者の ID を付与する．

なお，理想的な登場人物アノテーションの具体例を図 1 に示す．

登場人物アノテーションの自動化には，いくつかのバリエーションが考えられ，それぞれ実現の難易度が異なる．理想的な自動化は，小説テキストを与えると，上記の要素にタグを挿入したテキストが出力される完全自動化設定である．もう少し実現が容易な設定は，あらかじめ用意した登場人物のリストも入力として与える設定である．

完全に未知の小説に対しては，あらかじめ登場人物リストを準備することは非現実的である．一方，よく知られた小説では，ウィキペディア等に登場人物が紹介されていることも多く，登場人物リストの作成は，現実的である．

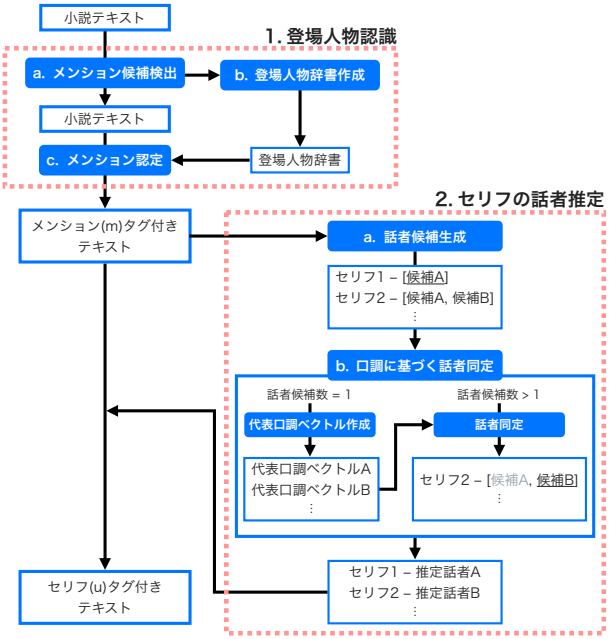

図 2　システム構成

3　システム

本研究では，登場人物アノテーションを完全自動化設定で実行するシステムを作成した．ただし，実行の途中経過で作成される登場人物辞書に対して，人手による介入 (追加・削除・修正) が可能である．これにより，すでに存在する登場人物リスト等が利用できる場合は，これも活用できる．

システムの構成を図 2 に示す．この図に示すように，システムは，大きく (1) 登場人物認識，(2) セリフの話者推定，の 2 つのモジュールから構成されている．

3.1　登場人物認識

登場人物認識では，小説に登場する人物を検出し，テキスト中のメンションに人物 ID を付与することを行う．登場人物認識モジュールには，すでに提案済の登場人物認識システム [1] を拡張したものを使用する．

この登場人物認識システムは，次のような 2 パス構成となっている．

1. 小説テキストを走査してメンション文字列を収集したのち，それらの文字列が同一人物を指し示すかの判定を行い，登場人物辞書を作成する．
2. 再度小説テキストを走査し，作成した登場人物辞書を用いて，メンションにタグを挿入する．

作成する登場人物辞書のエントリは，登場人物の ID と，その人物のメンション文字列集合である[1)]．すでに述べたように，第 2 パスを実行する前に，この辞書を確認し，必要ならば修正することも可能である．

提案済のシステムが対象とするメンション文字列は，固有表現 (名前・ニックネーム) のみであるが，本モジュールでは，一人称小説の地の文に現れる一人称表現も対象とする．具体的には，以下の手順で，一人称小説か否かの判定，一人称表現の決定，および，人物同定を行う．

1. 小説テキストの地の文に現れる一人称代名詞を種類別に収集する．
2. 最もよく出現する一人称代名詞の頻度 (出現数／地の文の数) が，ある閾値 F を超えた場合，その小説を一人称小説と認定し，その一人称代名詞をメンション文字列に含める．
3. 固有表現の出現比 (セリフ内出現数／地の文内出現数) が閾値 R を超える人物のうち，出現数が最も多い人物を，一人称代名詞の人物とする．

実験では，それぞれ $F = 0.1, R = 5$ を用いた．

3.2　セリフの話者推定

登場人物認識では，テキスト中のセリフに，その話者の情報 (人物 ID) を付与することを行う．セリフ話者推定モジュールには，石川らのシステム [3] を一部修正したものを用いる．

石川らの話者推定方法は，大きく 2 ステップに分けられる．

1. 話者候補生成
   いくつかのヒューリスティックを適用し，それぞれのセリフ対して話者候補リストを作成する．話者候補が 1 名の場合は，話者を確定させる．
2. 口調に基づく話者同定
   話者が確定しなかった 10 文字以上のセリフに対し，口調ベクトルの類似度を用いた方法で話者を決定する．具体的には，まず準備段階として，前者のステップで話者が確定したセリフを用いて，それぞれの登場人物に対する口調ベクトル (代表口調ベクトル) を作成する．各セリフ

1)　小説では，同一人物が複数の表記で記述されることは珍しくない．

の話者は，そのセリフを口調ベクトルに変換したのち，話者候補の代表口調ベクトルの中から，そのセリフの口調ベクトルに最も距離が近い代表口調ベクトルを求めることにより決定する．

本モジュールと石川らのシステムの違いは，以下のとおりである．

- 石川らのシステムの話者推定対象セリフは，主要人物のセリフのみである．これに対して，本モジュールは，すべてのセリフを対象とする．
- 石川らのシステムでは，主要登場人物のリスト (登場人物辞書に相当する情報) が入力として与えられ，話者候補の初期リストの作成 (セリフの周囲に出現する人物の検出) では，このリストに含まれる人物表記を文字列検索で収集する．これに対して本モジュールは，すでに付与されているメンションタグを検索し，話者候補の初期リストを作成する．なお，メンションタグは，登場人物認識で認定されたすべての人物に対して付与されるため，セリフがない登場人物にも付与されている．

4 実験

作成したシステムを用いて，登場人物アノテーションがどの程度自動化できるかを調べるための実験を行った．システムの入力は小説テキストであり，出力は m タグと u タグが付与されたテキストである．本実験では，表 1 に示す小説 3 作品を用いた．小説テキストは，OCR でテキスト化したのち，これを人手で修正したデータを用いた．

これらの作品のうち，作品 1 は，登場人物認識システム [1] の開発に用いた作品である．なお，この表には，登場人物認識の精度を示したが，この値は，登場人物認識システムの評価時の値である[2)]．

実験結果を表 2 に示す．この表には，本システムの実行結果に加え，比較のために，石川らの話者推定システム [3] を適用した場合の結果も示した．

先に述べたように，石川らのシステムでは，「正解の主要登場人物辞書」が与えられる．これに対して，本システムでは，登場人物認識を自動化し，自動推定された登場人物辞書を用いる．つまり，2 つのシステムの比較では，登場人物の情報の抽出を手動から自動化に変更した場合に，どのような差異が生じるかが興味の中心となる．なお，石川らのシステムでは，話者推定の第 1 ステップで，主要登場人物しか話者候補に含めないため，それ以外の人物のセリフの話者推定は，すべて誤りとなる．

なお，いずれのシステムでも，口調に基づく話者同定では，10 文字以上のセリフのみを推定対象とする．そのため，このステップに進んだ 10 文字未満のセリフは，話者を推定できず，話者未確定セリフとして残る．

表 2 の一番右端の精度 (赤字) が，すべてのセリフのうち，正しく話者を推定できたセリフの割合を表す．作品 1 では 66.2%であり，約 2/3 のセリフの話者を正しく決定できているが，作品 3 は 52.6%，作品 2 は 45.7%であり，精度は高くない．しかしながら，「正解の主要登場人物辞書」を与える石川らのシステムとの精度差は-2%程度であり，許容できる範囲である．

作品 2 や 3 の推定精度が低い原因のひとつは，これらの小説では 10 文字未満の短いセリフが多い (約 1/4) ためである．これらを除外した精度 (表の「話者確定の全体の精度」(青字)) は，3 作品とも 60%以上である．これらの精度は十分とは言えないが，小説の各セリフに話者を付与する作業の労力削減には寄与すると考えられる．

表 2 の「話者候補生成」は，話者推定の第 1 ステップで話者が確定したセリフの正誤数を，「口調」は，話者推定の第 2 ステップの話者同定結果の正誤数を示している．それぞれ主要人物のセリフとそれ以外の人物のセリフに分けて，値を示した．当然のことながら，話者候補に含める人物を主要人物に限定した場合，主要人物のセリフに対する推定精度は向上する．これは，各セリフに対する話者候補の初期リスト作成時の話者候補数 (表 3) が減少し，その結果，第 1 ステップで話者が確定する (代表口調ベクトルの作成に使用できる) セリフ数が増え，それが，代表口調ベクトルの質向上に寄与するからである．しかしながら，その代償として，それ以外の人物のセリフの話者推定がすべて間違うことになる．

一般に，主要人物のセリフ数と比較して，それ以外の人物のセリフ数は少ない．そのため，推定話者を主要登場人物に限定するという戦略も，一つの選択肢として存在する．本システムは，登場人物辞書に対する介入を許すため，この選択肢も実行可能である．ただし，登場人物を主要登場人物とそれ以外

2) 小説全体における，全登場人物のメンションの認識精度ではなく，一部の章・節に出現する主要登場人物のメンション (固有表現のみ) の認識精度である．

表 1　対象作品

番号	著者	タイトル	主要人物数	総話者数	物語の視点	登場人物認識の精度
1	米澤穂信	『氷菓』[2]	4	10	一人称	97.8%
2	谷川流	『涼宮ハルヒの憂鬱』[4]	5	14	一人称	100%
3	榎宮祐	『ノーゲーム・ノーライフ』[5]	4	15	三人称	90.5%

表 2　実験結果

作品	手法	人物	話者確定 全体 正	誤	精度	話者候補生成 正	誤	精度	口調 正	誤	精度	話者未確定	計	精度
1	石川ら	主要人物	677	135	83.4%	207	15	93.2%	470	120	79.7%	141	953	71.0%
		それ以外	0	49	0.0%	0	12	0.0%	0	37	0.0%	26	75	0.0%
		計	677	184	78.6%	207	27	88.5%	470	157	75.0%	167	1028	65.9%
	本研究	主要人物	648	157	80.5%	159	13	92.4%	489	144	77.3%	148	953	68.0%
		それ以外	33	28	54.1%	11	7	61.1%	22	21	51.2%	14	75	44.0%
		計	681	185	78.6%	170	20	89.5%	511	165	75.6%	162	1028	66.2%
	差分		+4	+1	±0.0%	-37	-7	+1.0%	+41	+8	+0.6%	-5		+0.3%
2	石川ら	主要人物	572	219	72.3%	200	25	88.9%	372	194	65.7%	270	1061	53.9%
		それ以外	0	104	0.0%	0	48	0.0%	0	56	0.0%	33	137	0.0%
		計	572	323	63.9%	200	73	73.3%	372	250	59.8%	303	1198	47.7%
	本研究	主要人物	519	267	66.0%	157	22	87.7%	362	245	59.6%	275	1061	48.9%
		それ以外	28	78	26.4%	6	6	50.0%	22	72	23.4%	31	137	20.4%
		計	547	345	61.3%	163	28	85.3%	384	317	54.8%	306	1198	45.7%
	差分		-25	+22	-2.6%	-37	-45	+12.0%	+12	+67	-5.0%	+3		-2.0%
3	石川ら	主要人物	705	236	74.9%	201	42	82.7%	504	194	72.2%	239	1180	59.7%
		それ以外	0	55	0.0%	0	15	0.0%	0	40	0.0%	52	107	0.0%
		計	705	291	70.8%	201	57	77.9%	504	234	68.3%	291	1287	54.8%
	本研究	主要人物	673	258	72.3%	155	30	83.8%	518	228	69.4%	249	1180	57.0%
		それ以外	4	57	6.6%	1	14	6.7%	3	43	6.5%	46	107	3.7%
		計	677	315	68.2%	156	44	78.0%	521	271	65.8%	295	1287	52.6%
	差分		-28	+24	-2.6%	-45	-13	+0.1%	+17	+37	-2.5%	+4		-2.2%

表 3　話者候補生成の初期リスト作成時の平均話者候補数

手法	作品 1	作品 2	作品 3
石川ら	3.00	3.07	2.28
本研究	3.71	3.50	2.66
差分	+0.71	+0.43	+0.38

の人物に分けることは，完全に未知の小説に対しては，ほぼ不可能であり，登場人物アノテーションシステムは，登場人物認識を含む完全自動化設定でも動作することが不可欠と考えられる．

5　関連研究

登場人物アノテーションは，登場人物認識とセリフの話者推定から構成されるが，これら 2 つのタスクは，これまで独立して研究されてきた．

登場人物認識は，出現検出と同一人物判定から構成される．出現検出には固有表現抽出の技術 (人名辞書や形態素辞書の利用 [6, 7]，動詞格フレームの利用 [8]，系列ラベリングの適用 [9]) が用いられる．同一人物判定には，ルール [6] や照応解析 [10] が用いられる．これらの研究の多くは後続タスク (人物情報抽出や登場人物間の関係抽出など) のための準備という意味合いが強いため，登場人物認識単体での出力形式がどのような形式であるかが明示されていない場合も多く，論文で提示された後続タスク以外の目的に利用できるか不明である．

セリフの話者推定は，表層的な手がかりのみを用いるもの [11, 12] と，口調を用いるもの [13, 3] がある．表層的な手がかりのみを用いるものは，日本語小説に特徴的な口調を利用していない．一方，後者の口調を用いるものであっても，性別のみを判別していたり，登場人物を人手で与えたりする．そのため，登場人物アノテーションの完全自動化には至っていない．

6　おわりに

本論文では，登場人物認識とセリフの話者推定を統合した登場人物アノテーションを説明したのち，それを実現するシステムを提示した．このシステムを用いて完全自動化設定で登場人物アノテーションを実行し，セリフの話者推定において，主要登場人物リストを与える先行研究と同程度の精度を得た．

参考文献

[1] 大島 一海, 窪田 智徳, 小川 浩平, 佐藤 理史. 固有表現を対象とした小説登場人物検出. 言語処理学会第 30 回年次大会発表論文集, pp. 728–733, 2024.

[2] 米澤 穂信. 氷菓. KADOKAWA, 2001.

[3] 石川 和樹, 小川 浩平, 佐藤 理史. 口調エンコーダを用いた小説発話の話者推定. 自然言語処理, Vol. 31, No. 3, pp. 894–934, 2024.

[4] 谷川 流. 涼宮ハルヒの憂鬱. 角川書店, 2003.

[5] 榎宮 祐. ノーゲーム・ノーライフ 1：ゲーマー兄妹がファンタジー世界を征服するそうです. KADOKAWA, 2012.

[6] 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.

[7] 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.

[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.

[9] Yuji Oka and Kazuaki Ando. Extraction of Novel Character Information from Synopses of Fantasy Novels in Japanese using Sequence Labeling. In Minh Le Nguyen, Mai Chi Luong, and Sanghoun Song, editors, **Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation**, pp. 505–513. Association for Computational Linguistics, 2020.

[10] 平川 大樹, 田村 直良. 物語テキストにおける登場人物の同一指示解析. 言語処理学会 第 22 回年次大会発表論文集, pp. 877–880, 2016.

[11] Du Yulong, 白井 清昭. 小説からの自由対話コーパスの自動構築. 言語処理学会第 25 回年次大会発表論文集, pp. 623–626, 2019.

[12] 銭本 友樹, 古俣 慎山, 宇津呂 武仁. BCCWJ を対象としたパターンマッチによる End-to-End 発話者分類. 言語処理学会第 29 回年次大会発表論文集, pp. 1995–2000, 2023.

[13] Yuki Zenimoto and Takehito Utsuro. Speaker Identification of Quotes in Japanese Novels based on Gender Classification Model by BERT. In Shirley Dita, Arlene Trillanes, and Rochelle Irene Lucas, editors, **Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation**, pp. 126–136. Association for Computational Linguistics, 2022.