物語テキストにおける登場人物の同一指示解析

平川 大樹　　田村 直良
横浜国立大学大学院 環境情報学府 情報メディア環境学専攻
hirakawa-daiki-yp@ynu.jp　　tam@ynu.ac.jp

1 はじめに

本研究では物語テキストにおける登場人物の同一指示解析について検討する.

今日, 音声合成技術の実用化により視覚障がい者が電子テキストにアクセスしやすい環境が整ってきた. 視覚障がい者への娯楽分野での支援を目的とし, 吉田らはラジオドラマ生成システムを提案した [1] [4]. これは物語テキストを入力として, 従来の音声合成器で生成された単一話者による単調な読み上げではなく, 複数話者による読み上げや話者の感情を考慮した発話, 環境を表現する効果音,BGM を含めることによって, より豊かな音声で表現するラジオドラマ作成のためのシステムである. 現時点では, アノテータが各種情報を人手で割り振っているため, それらの自動化が求められている.

物語テキスト中の発話文の話者同定や発話の感情推定のためには, どのような人物が登場するのか, 発話の際に登場人物がどのような感情なのか, 登場人物にはどのような特徴があるのかなど登場人物の情報抽出が必要となる. しかし, 物語テキスト中では同一の登場人物でも固有名詞や普通名詞, 代名詞と様々な形で表出される. そこで, 本研究では, 物語テキスト中の人を表す表現がどの登場人物を指しているか解析する手法について検討することを目的とする.

関連分野の研究としては, 物語の話者同定や照応解析の問題がある. 物語の話者同定について, 神代らは, 物語テキスト中の発話文の発話者がどの文に存在するか, という手法で話者同定を行った [2]. この研究ではある発話文における発話者と異なる文における発話者は同一人物か, といったテキスト中の表現が指示する対象の同一性までは考慮していない. 飯田らは照応詞候補に対して最も先行詞らしい候補（最尤先行詞) を同定した後, その対が照応関係にあるか否か判定を行うという手法で, 名詞の指示性を考慮せずに照応関係にない対を棄却する解析モデルを提案した [3]. この研究は新聞社説において人を表す表現の同一指示性や政策を表す表現の同一指示性などを包括的に解析している. しかし新聞社説では代名詞の表出回数が少ないため代名詞に関する精度が低く, 代名詞とそれ以外に関する表現では解析手法を分ける必要があると考えられる.

我々の手法では, まず物語テキストが記述する世界を設定し, テキスト中の人を表す表現に対して「指示対象」を生成する. 次にそれぞれの指示対象について同一性を判断し, 同一の対象を統合するという手続きで同一指示関係を解析する. 指示対象の統合にはルールによる統合, 照応関係の観点による統合, 指示対象の素性を手がかりにした統合という 3 つの手法を組み合わせる.

以下本稿では, 第 2 章で表現の指示対象の同一性について述べ, 第 3 章で登場人物の同一指示解析手法について述べる. 第 4 章では登場人物の同一指示解析の実験を行い, 実験結果について総括する. 第 5 章でまとめと今後の課題について述べる.

2 表現の指示対象の同一性について

2.1 記述世界

物語テキストには登場人物の行動や心境, 場面の情景などが描かれており, その世界の中でストーリーが展開していく. このように物語テキストが記述する世界を記述世界と呼ぶ. 人を表す表現の指示対象は記述世界中の人物である.

2.2 名詞の指示性

- 固有名詞とその指示対象
  記述世界中の登場人物の名前であり, その人物を指示対象として持つ.
  (例)メロス は激怒した.

- 普通名詞とその指示対象
  記述世界中の登場人物の通称や職業である. 指示対象は記述世界中の人物である.
  (例) メロスは, 村の 牧人 である.
- 代名詞・ゼロ代名詞とその指示対象
  物語テキスト中で固有名詞, 普通名詞の代わりに人を表す表現で, 性別の情報を含んでいる場合もある. 指示対象は記述世界中の人物である.
  (例) たちまち 彼 は, 巡邏の警吏に捕縛された.
  (例)「市を暴君の手から (私が) 救うのだ.」とメロスは悪びれずに答えた.

2.3 同一性について

- 結束構造の観点による指示性の解析
  照応関係にある 2 つの表現の指示対象は同一である事に加え, 主題が人物である場合の主題連鎖関係 (結束性) は同一である.
- 推論による指示性
  同一の指示対象かどうか知るために人間が持っている一般的な知識を必要とする場合がある.

  (例文 1) この 妹 は, 村の或る律気な一牧人を, 近々, 花婿として迎える事になっていた.
  (例文 2) メロスは, それゆえ,花嫁 の衣裳やら祝宴の御馳走やらを買いに, はるばる市にやって来たのだ.

  この 2 つの文の (例文 1) の妹が結婚するという情報から,(例文 2) の花嫁が妹であるということを推論している.

3 登場人物の同一指示解析

以下の手順で登場人物の同一指示解析を行う.

(1) 指示対象の作成

(2) 指示対象の同一化

  (a) ルールに基づいた統合

  (b) 照応関係に基づいた統合

  (c) 指示対象の素性を手がかりにした統合

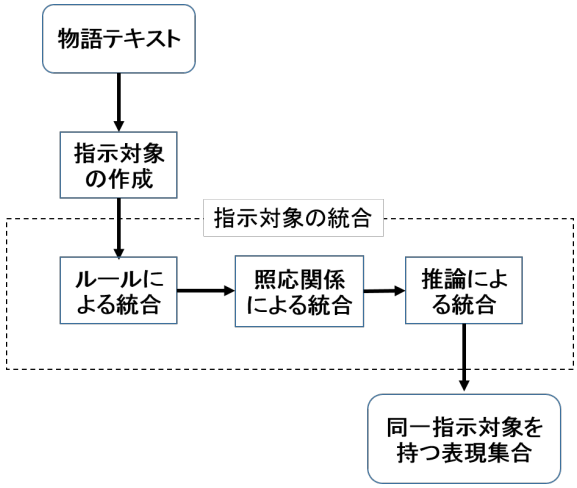

図 1: 登場人物同一指示解析手法の概要

3.1 指示対象の作成

人を表す表現から指示対象として認識し, 同時に以下の情報を抽出し, 指示対象の素性としてまとめる.

- 名前
  固有名詞をその人物の素性 (名前) とする. 実験で扱ったテキストでは事前に人手で固有名詞の情報を与えた.
- 通称
  職業ではない普通名詞をその人物の素性 (通称) とする.
- 職業
  日本語語彙体系において職業を表す表現を, その人物の素性 (職業) とする.
- 性別
  日本語語彙体系において性の情報を含んだ表現を, その人物の素性 (性別) とする.
- 年齢
  日本語語彙体系において年齢の情報を含んだ表現を, その人物の素性 (年齢) とする. 年齢は幼児, 少年・少女, 成人, 老人の 4 つに分類される.

情報が抽出できない場合, 不明とする.

3.2 ルールに基づいた統合

以下のルールに従って, 指示対象を統合する.

- 素性 (名前) が一致する表現の指示対象を統合.
- テキストにおける複合名詞が
  < 人を表す表現 1 >< 人を表す表現 2 >
  の場合,< 人を表す表現 1> と < 人を表す表現 2>
  の指示対象を統合.

- テキスト中に
  < 人を表す表現 1 >< 記号 >< 人を表す表現 2 > が表出する場合,< 人を表す表現 1> と < 人を表す表現 2> の指示対象を統合.

このルールにより固有名詞が一致する表現や名前と職業・通称が複合名詞として表出する表現などの指示対象が統合される.

3.3 照応関係に基づいた統合

代名詞のみを照応詞として, 飯田らの手法を用いる. 飯田らの手法は以下の手順で行われる. なお, ゼロ代名詞照応は現在実装されていない.

3.3.1 最尤先行詞候補の同定

最尤先行詞候補の同定ではトーナメントモデルを用いて照応詞候補の前方にある尤も先行詞らしい候補を最尤先行詞とする. まず, 先行詞候補群から照応詞候補に最も近い 2 つの候補から, 学習にされた判定器により一方を選び, 選ばれなかった先行詞候補を先行詞候補群から削除する. これを繰り返すことにより最尤先行詞を同定する. 本研究では, 照応詞候補の文前方 5 つの文に含まれる先行詞候補を対象とした.
照応詞と先行詞, 最も照応詞に近い先行詞ではない人を表す表現の 3 つの表現の組を訓練事例とした.

3.3.2 照応詞関係の認定

照応関係の認定では,3.3.1 節で得られた照応詞候補と最尤先行詞候補の対が真に照応関係にあるか否かを判定する. 非照応詞と最尤先行詞の対を負例として学習した判定器により, 照応関係にない対を棄却する事が可能となる.
3.3.1 節の手法で得られた照応関係にある対を正例, 照応関係にない対を負例として訓練した.

3.3.3 素性

表 1 は 3.3.1 節の判定器と 3.3.2 節の判定器で用いた素性である. 最尤先行詞同定では照応詞候補と 2 つの先行詞候補から, 照応関係の認定では照応詞候補と最尤先行詞候補から素性を抽出する. ((∗) は最尤先行詞同定でのみ用いる素性.)

表 1: 実験に用いた素性

素性
照応詞候補 (先行詞候補) があ系の代名詞であるか
照応詞候補 (先行詞候補) がそ系の代名詞であるか
照応詞候補 (先行詞候補) がこ系の代名詞であるか
先行詞候補が固有名詞であるか
照応詞候補 (先行詞候補) に続く助詞
照応詞候補 (先行詞候補) が日本語語彙体系の人を表す名詞かどうか
照応詞候補 (先行詞候補) の性別
照応詞候補 (先行詞候補) が年齢
照応詞候補とそれぞれの先行詞候補の距離
先行詞候補間の距離 (∗)
照応詞候補 (先行詞候補) が文頭であるか
照応詞候補 (先行詞候補) が文末であるか

3.4 指示対象の素性を手がかりにした統合

指示対象の素性を手がかりにした統合では, 隣り合う文にある 2 つの人を表す表現に着目し, 指示対象の統合を行う. 隣り合う文にある 2 つの人を表す表現が次に示す (1) を満たしながら,(2) を満たさないとき, その 2 つの指示対象を統合する.

(1) 指示対象の素性 (職業または通称) が等しい
(2) 指示対象の素性 (性別または年齢) が異なる

ただし, 指示対象の素性 (性別または年齢) が不明のとき,(2) は常に満たさないとする.

4 実験と評価

青空文庫に収録されている物語から 3 編を対象とし, 登場人物の同一指示解析を行った. 表 2 にそれぞれの物語の情報, 正解として作成した登場人物数を示す. 同

表 2: 実験に使用した物語

タイトル	走れメロス	羅生門	白銀の失踪
著者名	太宰治	芥川龍之介	コナン・ドイル
登場人物数	5	2	7

一の指示対象を持つ表現の集合を同一人物指示集合と呼ぶ. 付録の評価実験では人手で作成した人を表す代名詞とその先行詞のペアを正解とし,3.3 節の手法で生成された結果と比較し, 適合率, 再現率で評価した.4.1 節では, テキスト中の人を表す表現が誰を指しているか人手で正解を作成し, 本研究の手法から作成された同一人物指示集合と比較して適合率, 再現率で評価した. このとき, 作成された同一人物指示集合の指示対象が持つ名前の人物と比較をした. 名前が不明の際には再頻出する通称, 職業を持つ人物との評価を行った.

4.1 同一人物指示集合の評価

3 章で説明した一連の手法から作成された同一人物指示集合と人手で作成した同一人物指示集合を比較した. 作成された同一人物指示集合の中で, 同一の人物を指示しているが集合が分かれている場合は最も要素数が大きい集合と正解を比較した. 作成された同一人物

表 3: 各物語における同一人物指示集合の評価

タイトル	適合率	再現率
走れメロス	1	0.264
羅生門	1	0.348
白銀の失踪	0.919	0.245

指示集合の多くが, 名前・職業・通称のみから成り立っていたため適合率が高い. 次に再現率低下の原因を考察する.

- 照応解析誤り
  最尤先行詞同定の誤りにより, 正しい照応詞と先行詞の対を抽出できなかった. このため, 後の同一人物指示集合にも代名詞がほとんど包含されず再現率が低下した. また先行詞が代名詞となる事もあり, その代名詞がその代名詞の指示対象が解析できなければ, どちらも同一人物集合に包含されないため, 再現率が低下する.
- テキスト中に離れて出現する表現の統合ができない
  『羅生門』では下人と表現される人物を示す集合が 7 つ作成された. これは下人という表現が離れて表出したため, 異なる集合として作成された. 通称や職業の一致をテキストを通して行えば集合は統合できるが,『走れメロス』における「石工」や「牧人」のように異なる人物でも同一の職業を持つ場合に誤りが発生する事が考えられる.

その他の誤りの例としては同じ固有名詞で表現されていても別人である場合があった.

- 『白銀の失踪』における「ストレーカさん」は「ストレーカ夫人」を意味しているが, 夫も「ストレーカ」と呼ばれている.

5 終わりに

物語テキストにおける登場人物の同一指示解析手法を提案した. 実験全体での適合率は 0.973、再現率は 0.286 であった. 再現率を上げるために, 照応解析の改善や同一人物指示集合の統合方法を改善する事が課題である. 加えて, ゼロ照応, 主題解析を取り入れ, 精度向上を図りたい.

参考文献

[1] 吉田有里, 奥平康弘, 田村直良.E-051 音声合成による朗読システムに関する研究 (自然言語・音声・音楽, 一般論文). 情報科学技術フォーラム講演論文集,Vol.8,No.2,pp.377-380,aug 2009.

[2] 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.2008 年,3 月,pp.380-383.

[3] 飯田龍, 乾健太郎, 松本裕治, 関根聡, 最尤先行詞候補を用いた日本語名詞句同一指示解析. 情報処理学会論文誌,Vol.46,No.3.Mar,2005.pp.831-844

[4] 金子つばさ, ラジオドラマ生成システムにおける読み上げデータの作成支援, 横浜国立大学, 卒業論文

付録

照応解析の評価

照応関係の認定では leave-one-out 交差検証を行い評価した.

表 4: 照応解析評価実験

適合率	0.923
再現率	0.063

最尤先行詞候補同定での誤りが多く, 再現率が低下した. 発話文と地の文の扱いを考慮しなかった事, 一人称, 二人称, 三人称の扱いを考慮しなかったため, 最尤先行詞が正しく同定されなかったと考えられる.