述語情報を利用した小説の登場人物の抽出

米田 崇明　篠崎 隆宏　堀内 靖雄　黒岩 眞吾
千葉大学大学院 融合科学研究科

1　はじめに

電子化に伴うコストの削減により，小説などの書籍はインターネットを通じて容易に入手できるようになった．これによりユーザの選択幅を広げることになるが，膨大な選択肢の中で自分の嗜好に合うものを探すことは難しい．このためユーザが読む書籍を容易に探す手掛かりとなる情報が必要となる．

書籍の中でも，小説を選択するための情報を作成することは人手に頼ることが多い．これは情報を作成する際には，内容を理解している必要があるからである．しかし，多くの小説が発行されている中で人手に頼ることは金銭面や時間面において不利である．このため，自動的に小説を解析する技術の確立が急がれる．

小説を解析する上で重要な要素の一つとして登場人物がある．登場人物に関連する研究は，例えば縣らの研究 [1] のように登場人物を手動で決定していることが多い．手動で登場人物を定めることは確実だが，作品毎に登場人物を人手で抽出しなくてはならない以上，汎用性に欠ける．そこで，小説の登場人物を自動抽出する手法が考案されている．例えば馬場らの研究 [2] では，形態素解析の辞書に人名を追加することによって登場人物名を抽出している．しかし，辞書に依存する手法は未知語が多数出現する人物名においては頑健な手法とは言えない．そこで本稿では，未知の人物名に対しても動作するよう各単語の主語としての出現頻度と述語の性質から登場人物を推定する手法を提案する．提案手法は，人物候補を抽出した後それが登場人物であるか判定する，2 段階のプロセスに基づいている．

以下では，2 章及び 3 章において人物候補の抽出及び登場人物の判別についてそれぞれ説明する．その評価のための実験を 4 章にて行い，最後に 5 章でまとめを行う．

2　人物候補の抽出

本研究では，“ 小説内において登場人物は主語として必ず出現する ”という仮説を立てた．小説は登場人物の行動や言動が中心となって物語が進むことが多いからである．この仮説に基づき，小説中の各文の主語を登場人物の候補 (以下，人物候補とする) として抽出した．

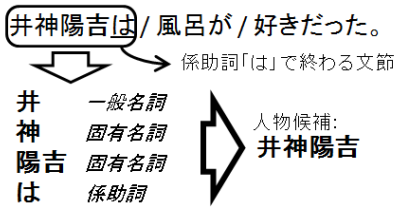

図 1: 人物候補 (主語) の抽出

文中の主語は，句読点などを除くと係助詞『は』や格助詞『が』で終わる文節に含まれていることが多い．実際に，このことを利用して登場人物の抽出を試みた例もある [3]．これに加えて,『A と B は (が)〜』のように，並立助詞『と』が含まれる文節の後に『〜は』『〜が』の文節が続く文の場合は,『〜と』の文節に含まれる語群も主語と言える．

人物候補を抽出する際には図 1 のように形態素及び文節の情報を利用した．構文解析を行う際に文節も区切られるため文節の情報は構文解析の結果を用いた．その上で『〜は (係助詞)』『〜が (格助詞)』『〜と (並立助詞)』となる文節を探索し末尾にある助詞などを除いた形態素群を主語とみなし人物候補として抽出した．

3　登場人物の判別手法

抽出された人物候補の中には登場人物も含まれるが登場人物ではないものも含まれている．このため登場人物かどうか判別する必要がある．ここでは，以下の二つの手法を用いて登場人物を判別することを検討した．

- 人物候補の主語としての局所出現頻度を利用した手法
- 人物候補と述語情報の関係を利用した手法

また，両者を統合した手法についても検討した．

3.1 前処理

前述のように小説は登場人物の行動や言動が中心となって物語が進むことが多いことから，複数回出現した主語は登場人物名である可能性が高い．

人物候補 s が主語として出現した回数を考える．ただし，松本らの研究 [4] によれば話し言葉は形態素解析の精度が低下することから人物の検出精度も低下させると予想されるため，小説内での発言 (鍵括弧で判断) である文は出現回数から除外した．また，形態素解析結果に『接尾』とつけられた形態素をもつ人物候補は,『接尾』を外した状態の主語と同一視した．例えば『A』と『Aさん』は同じ人物候補としてカウントされる．さらに，主語が検出されない文は会話文を除いた直前の文に主語があればその文と同じ主語とみなし主語を補った．

主語として出現した回数を利用すれば主要な登場人物は抽出できると考えられる．そこで，出現回数が少ない人物候補 (本研究では 3 回未満) は登場人物ではないとみなして人物候補から削除した．また，抽出された主語の中には『これ』『それ』など登場人物を指さない語も含まれている．このため人物になり得ない語はストップワードとして除外されるようにした．また，平仮名・片仮名一文字の名前も人物として相応しくないため除外した．

3.2 局所出現頻度による判別

前処理で残った人物候補の中で，出現する場面が限られ出現回数が多くない登場人物は同じくらいの出現回数である登場人物でない語と区別できない可能性がある．しかし，登場人物は一度出現すればしばらく出現する傾向にある．その点に着目して一定区間ごとの局所出現頻度を測定し，それに従って主語を分類した．

提案手法では，人物候補の局所出現頻度を図 2 のように小説中の一定文数 (これを窓と呼ぶ) における特定の主語の出現回数に基づいて算出する．窓は，特定の短い場面に登場する人物を検出するための狭い窓 R_1 と，文書全体の局所出現頻度が高い人物を検出するための広い窓 R_2 の二種類を設定した．これを利用してある区間の局所出現頻度が高い人物候補を登場人物と判断することによって，出現回数が多くない人物候補も登場人物と判別できると期待される．

窓の中心を現在位置 n として R_1,R_2 それぞれの窓幅の範囲を移動していき，各窓内の文に対象の人物候補 s が出現するかどうか調べる．その文の主語が対象の人物候補の場合は $a_{sm}=1$ となり，そうでない場合は $a_{sm}=0$ となる．窓の中心 n での人物候補 s の出現率を窓 R_1,R_2 を利用してそれぞれ f_{R_1sn},f_{R_2sn} とした．その値は以下のように計算できる．

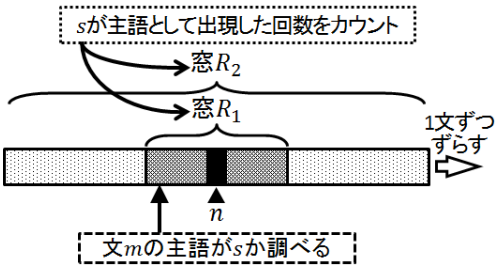

図 2: 窓を利用した局所出現頻度の算出

$$f_{R_isn}=\sum_{m=n-r_1/2}^{n+r_1/2}\frac{a_{sm}}{r_i} \quad ,(r=1,2)$$

これを全ての文に対して求め，その集合を F_{R_1sn},F_{R_2sn} とする．このうち注目すべき点は F_{R_1sn},F_{R_2sn} の最大値 $F^{max}_{R_1sn}$,$F^{max}_{R_2sn}$ である．これが高い値を示しているなら，全体の出現回数が少なくても人物候補 s は人物である可能性が高いと言える．このため人物候補 s を人物と判断する際は $F^{max}_{R_1sn}$,$F^{max}_{R_2sn}$ のみを考慮した．

これらを用いて登場人物を判別する際は線形判別分析 (LDA) を利用した．つまり，判別関数を

$$l_s=\alpha_1F^{max}_{R_1sn}+\alpha_2F^{max}_{R_2sn}+\beta$$

とし，LDA によって α_1,α_2,β の値を推定した．

3.3 述語情報による判別

局所出現頻度を利用した登場人物の判別手法は，当然その人物候補が人物かということは区別していない．つまり以下の状況が発生しうる．

- 登場人物でない語が連続的に主語になり登場人物と判断
- 時々出現し出現回数も少ない登場人物を登場人物でないと判断

これを解決する手法として，人物候補が人物らしい語であるか判断することが考えられる．そこで注目したのが主語に対応する述語である．述語に含まれる品詞として考えられるのは動詞・形容詞とサ変動詞 (〜する) に接続する名詞、形容動詞語幹の名詞である．述語の中には『言う』『話す』など人物が主語になりやすいものと『始まる』『過ぎる』など人物になりにくいものがある．そこでいくつかの小説を用いて主語と述語の共起関係を学習することにより登場人物かどうかを判別する手法を提案する．

まず学習に用いる小説で 3 回以上出現する人物候補を自動で抜き出し，人物候補 (主語) と共起する述語

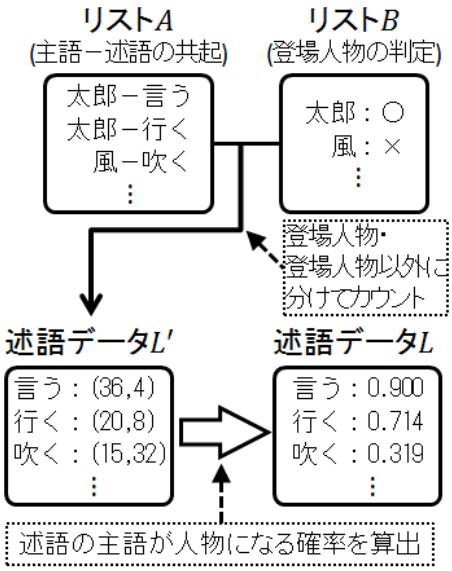

図 3: 述語情報の学習の流れ

を抜き出してリストにする (リスト $\boldsymbol{A}$)．ただし述語は全て原形を抜き出すものとし,『する』などのサ変動詞は除外した．また，主語を補完した文は述語を含むと判断できる最初の文節を述語として抜き出した．さらに，人物候補が人物であるかそうでないかの分類が記述されたリストを作成する (リスト $\boldsymbol{B}$)．これは主語の抽出は自動，それが人物かどうかは手動で判断したものである．

作成した 2 つのリストを利用して，図 3 のように述語データ $\boldsymbol{L}$ の作成を行う．リスト $\boldsymbol{A}$ からリスト $\boldsymbol{B}$ の要素を参照し，同じ主語のものを探す．リスト $\boldsymbol{B}$ で登場人物となっている主語は登場人物 q_1，そうでない語は登場人物以外 q_2 としてカウントして述語データ $\boldsymbol{L'}$ としてまとめる．述語データ $\boldsymbol{L'}$ は，述語 w と q_1,q_2 が各要素に含まれている．これを利用して述語 w の主語が登場人物になる確率 p_w が計算でき，

$$p_w = \frac{q_1}{q_1 + q_2}$$

となる．述語データ $\boldsymbol{L}$ は，この p_w と述語 w を要素としている．

述語データによって着目する述語の主語が人物になりやすさが得られる．これはその述語がどれだけ人物と係り受け関係にあったかを示しており，これが高い主語ほど登場人物である可能性が高い．この述語データを利用して人物候補から登場人物をそうでないものを判別する．本文中に 3 回以上出現する主語 s に対して以下の式を利用して s が登場人物に属する確率 $p_s(c)$ を求める．

$$p_s = \frac{1}{n}\sum_{i=0}^{n-1} p_{w_i}$$

ここで n は主語 s の出現回数である．述語が存在する全ての文を対象に人物のなりやすさ p_{w_i} を学習データ $\boldsymbol{L}$ から参照し，その平均値を出すことで主語 s が登場人物らしいかを求める．この値 p_s の大小により人物候補が登場人物かどうかを判断する．

表 1: 実験に使用した小説

母を尋ねて三千里	国貞えがく
セロ弾きのゴーシュ	冬の日
花のき村と盗人たち	馬の脚
野のはくちょう	クララの出家
ある男の堕落	電気風呂の怪死事件
気狂い機関車	広重と川獺
判決	アラジンとふしぎなランプ
犬神娘	橇
能面の秘密	黄金鳥
ヴィヨンの妻	一兵卒
犬を連れた奥さん	自転車嬢の危難
浅間噴火口	無月物語
アッシャー家の崩壊	聖家族
チャアリイは何処にいる	鏡中の美女
道づれ	かのように

3.4 両者を用いた登場人物の判別

3.1 で提案した局所出現頻度を利用した手法と 3.2 で提案した述語情報を利用した手法を合わせて人物候補から登場人物を判別する手法として，LDA を利用した判別手法を提案する．具体的には，3.1 で利用した判別関数

$$l_s = \alpha_1 F^{max}_{R_1 sn} + \alpha_2 F^{max}_{R_2 sn} + \beta$$

に主語 s と共起する述語により人物らしいかを計算した $p_s(c)$ を加えて

$$l_s = \alpha_1 F^{max}_{R_1 sn} + \alpha_2 F^{max}_{R_2 sn} + \alpha_3 p_s(c) + \beta$$

とした式の α_1,α_2,α_3,β を LDA によって推定する．

4 評価実験

4.1 実験条件

提案手法の検証のため，青空文庫 [5] より表 1 の短編小説 30 作品を選出し人物候補から登場人物を判別する実験を行った．これらには，ストップワードに含まれず 3 回以上詩出現した人物候補となるものが 672 個あり，そのうち 320 個は登場人物である．人物抽出に必要な情報を得るため，形態素解析には MeCab [6] を，構文解析には CaboCha [7] を利用した．実験では 1 作品をテストデータ，29 作品を訓練データとして用いるクロスバリデーションで検証を行った．本手法では登場人物の抽出を目的としているため，抽出された登場人物に対する適合率と再現率，それらを基にした F 値を評価指標としている．なお，述語を用いた判

表 2: 局所出現頻度による判別結果

	適合率	再現率	F 値
closed	84.1 %	59.4 %	69.6 %
open	82.0 %	59.8 %	67.1 %

表 3: 述語情報による判別結果

	適合率	再現率	F 値
closed	88.5 %	95.1 %	91.7 %
open	60.0 %	84.6 %	68.8 %

別手法を用いる場合は $p_s(c)$ がどれだけあれば登場人物と判定するかを示す閾値を推定する必要がある．訓練データを用いて閾値を 0.01 から 0.99 まで 0.01 ずつ変化させていき，F 値が最大値を示した閾値を用いて open テストを行い登場人物かどうか判定した．

4.2 実験結果

4.2.1 局所出現頻度による判別

人物候補の主語としての局所出現頻度を用いた手法の実験結果を表 2 に示す．表 2 から，この手法では closed で 69.6%，open で 67.1%と性能があまり変わらなかった．これは登場人物の局所出現頻度は作品によってあまり変化しない，若しくは似た特徴を持つ作品が多いことが要因と考えられる．また適合率はやや高いものの，再現率は低く全体の性能を引き下げている．性能の誤判別が起こる人物候補の特徴としては，出現回数が少ないことが挙げられる．本手法では出現回数を考慮せずに判別を行ったため，出現回数が多い人物候補による高い局所出現頻度が出現回数の少ない人物候補に影響したことが原因と考えられる．

4.2.2 述語情報による判別

主語と述語の共起情報を用いた手法で登場人物の判別を行った結果を，表 3 に示す．closed では F 値が 91.7%と良い性能が出ているものの open になると F 値が 68.8%と性能がかなり低下してしまった．これは，小説によって出現する述語の傾向 (言い回しや表記) が異なるため open では学習データに存在しない述語が多く，正しい結果が得られなかったのが原因であった．ただし，局所出現頻度と比べて F 値が closed で 22.1%，open でも 1.7%高くなっていることから，局所出現頻度を用いた方法よりも有効な手法であると言える．また，局所出現頻度を用いた手法とは逆に適合率は低く再現率が高い結果となった．適合率を改善するためには，学習する小説の数を増やす必要がある．

表 4: 頻度・述語による判別結果

	適合率	再現率	F 値
closed	83.1 %	97.3 %	89.6 %
open	60.3 %	91.9 %	71.5 %

学習する小説の数が多ければ，より正確な主語と述語の共起情報を得られ精度も上昇すると考えられる．

4.2.3 頻度・述語を用いた判別

局所出現頻度と述語の情報を両方用いた手法での結果を，表 4 に示す．この手法でも closed では 89.6%と性能が良かったが open では 71.5%と性能の低下が目立つ．これは適合率は述語情報での値が低いことによる影響と推測でき，述語情報のみを利用した手法と同様に学習する小説の数を増やせば精度も上がると考えられる．また，他の 2 手法と比べて F 値が高いことから，2 つの手法を組み合わせることによって良い判別性能が得られたと言える．

5 おわりに

本稿では，述語情報と局所出現頻度を用いることにより小説の登場人物を未知の人物名に対しても抽出できることを示した．今後の課題として，SVM 等の他の判別手法を用いる，小説の数を増やすなど精度を改善する手法の検討が挙げられる．また，抽出した登場人物から小説内の他の情報を抽出する手法の検討も必要である．

参考文献

[1] 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, Dec. 2010.

[2] 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, pp. 574–577, Mar. 2007.

[3] 小林聡. 場・時・人に着目した物語のシーン分割手法. 情報処理学会研究報告, Vol. 47, pp. 25–30, May 2005.

[4] 松本裕治, 伝康晴. 話し言葉の形態素解析. 自然言語処理研究会報告, Vol. 54, pp. 49–54, May 2001.

[5] 青空文庫, 1997. http://www.aozora.gr.jp/.

[6] 工藤拓. Mecab : Yet another part-of-speech and morphological analyzer, 2005.

[7] 工藤拓, 松本裕治. Cabocha: Yet another japanese dependency structure analyzer, 2001.