小説テキストを対象とした人物情報の抽出と体系化

馬場 こづえ　　藤井 敦
筑波大学大学院図書館情報メディア研究科
{kou,fujii}@slis.tsukuba.ac.jp

1 はじめに

インターネットや大容量電子媒体の普及によって，電子化テキストは身近になった．大量の電子化テキスト集合から目的のテキストを効率的に取得するためには，情報検索や自然言語処理などの研究分野で提案された技術が有用である．これらの分野で中心的に研究される対象は，新聞，論文，特許など，事実を客観的に伝えることを主たる目的とした「情報伝達テキスト」である．

しかし，電子化テキストには，小説，エッセイ，日記などのように，読んで楽しむことを指向したテキストや創作物としての価値を追求したテキストもある．筆者らは，これらを「娯楽・芸術テキスト」と総称し，中でも小説に焦点を当てた研究を行っている [1, 2].

テキストの内容で検索や分類を行う場合，情報伝達テキストは主にトピック（主題）を表す索引語（「情報検索」や「青色発光ダイオード」など）によって計算機上でモデル化される．

一方，小説テキストはストーリーや登場人物で検索されたり，分類されることがある．例えば「『ハッピーエンド』の話が読みたい」や「『頭脳明晰な探偵が主人公が登場する話』が読みたい」といった要求がある．しかし，索引語によるモデル化ではストーリーや登場人物の属性に基づく検索や分類には限界がある．

娯楽・芸術テキストを対象としたモデル化の研究には，因果関係によるモデル化 [3, 4] やシナリオ形式によるモデル化 [5] がある．

本研究は，登場人物に基づいて小説テキストをモデル化する．具体的には，英米文学の推理小説を対象に，テキストから自動的に人物相関図を作成する．英米文学に限定した理由は，外国人名がカタカナで表記されるので，カタカナ以外の語を排除することで人名抽出の精度を上げることができるからである．

2 提案する手法の概要

図 1 に本手法の概要を示す．入力は小説テキストで，出力は人物相関図である．長方形は処理を表し，円柱は使用する規則や辞書などの資源を表す．本手法は大きく分けて人物情報の「抽出」と「体系化」からなる．

「抽出」では，各処理を文単位で行うために，まずテキストを一文ずつに分割する．このときに会話文と地の文も分割する．次に，形態素解析結果に基づいて人名を抽出する．さらに，辞書と抽出規則を用いて人名の周辺文脈から人物の属性を抽出し，人名と属性をまとめてリスト化する．本研究では人物とその属性をまとめて「人物情報」と呼ぶ．

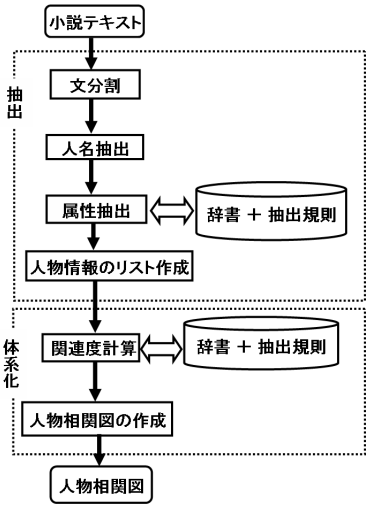

図 1: 本手法の概要

「体系化」では，場面における共起頻度を用いて特定二者間の関連度を計算する．場面に登場する人物を適切に特定するために，辞書と抽出規則を用いる．最後に，関連度に基づいて人物相関図を描画する．

3 節と 4 節で人物情報の「抽出」と「体系化」についてそれぞれ説明する．

3 人物情報の抽出

3.1 文分割

一般的に，人物の会話文は，かぎ括弧で括られて地の文と区別され，文末は句点で示される．そこで，句点やかぎ括弧を文末の根拠とし，テキストを文単位に分割する．

3.2 人名抽出

「グレゴリ 警部と ロス 大佐が居間で待っていた」という文から「グレゴリ」と「ロス」を人名として抽出するために，形態素解析を利用する．本研究では ChaSen[1] を使用し，品詞が「名詞-固有名詞-人名-一般」,「名詞-固有名詞-人名-姓」,「名詞-固有名詞-人名-名」と解析された形態素を人名として抽出する．

人名抽出の網羅性を高めるため,「8 万人西洋人名よみ方綴り方辞典 [6]」から人名を収集して ChaSen の辞書

[1] http://chasen.naist.jp/hiki/ChaSen/

に追加した．その結果，辞書中の人名見出しは 31685 件から 79985 件に増加した．

3.3 属性抽出

3.2 節で抽出した人物名の周辺文脈から，人手で作成した規則と表層一致した表現を人物の特徴として抽出する．抽出する属性の種類は「性別」,「年齢」,「年代」,「職業」,「身体的特徴」,「性格」である．ある属性に対して値が一つも抽出できなかった場合は「不明」とする．

抽出規則として，筆者らの研究 [1, 2] で作成した規則を用いる．今回は，抽出精度を高めるために，辞書の増強と係り受け解析の導入を行った．以下，6 種類の属性に関する規則とその利用法について個々に説明する．

3.3.1 性別

性別には「男性」と「女性」の値がある．性別がわかる語（男性，父，叔父など）や性別固有の一人称が含まれていれば，該当する値を付与する．同一の人物に対して，女性を表す語と男性を表す語の両方が出現している場合は，テキスト全体で多く出現している値を付与する．

3.3.2 年齢

登場人物の年齢表記を値として付与する．「17 歳」や「三十五才」といった表記を規則によって抽出する．数表記の次に「才」または「歳」の字が現れた場合を年齢表記と定義する．数表記には半角数字，全角数字，漢字表記があるので全ての表記に対応する．同一の人物に対して複数の年齢表記が出現している場合は，全ての表記を値として付与する．

3.3.3 年代

人間の一生を「乳幼児期」,「少年期」,「青年期」,「中年期」,「老年期」に区分し，それぞれを値とする．各年代を示す語をリスト化し，リスト中の語と表層一致した場合にその値を付与する．同一人物に対して複数の年代を示す語が出現した場合は，テキスト全体で多く出現している値を付与する．

なお，年齢と年代の対応関係は判断に個人差が生じるため，年齢から年代を判定することはしない．

3.3.4 職業

職業とその職業を示す語リストを「世界樹の下[2]」の職業による検索項目と,『角川類語新辞典 [7]』の「生産的職業」と「サービス的職業」の項目を参考に作成した．「世界樹の下」では，登場人物の属性（性別，身体的特徴，職業，性格など）で小説を検索することができる．

リスト中の語と表層一致した場合，職業名の値を付与する．複数の値が一致した場合は全て付与する．

3.3.5 身体的特徴

髪や瞳の色，声，体格など，容姿に関する特徴を値として抽出する．抽出する特徴の数に制限はなく，抽出した特徴をそのまま全て登場人物に付与する.「青い目」という特徴が規則から抽出された場合，それを値として付与する．

抽出規則は三通りあり，Perl の正規表現に準拠して表記すると以下のようになる．

- 「身体を表す語」(が | は)({ 形容詞 } | { 名詞 })
- { 形容詞 } (「身体を表す語」)
- { 名詞 } の（「身体を表す語」）

「身体を表す語」のリストを作成するために，既存の辞書 [8, 9] を参考にして 63 語を収集した．{ 形容詞 } は形態素解析結果で形容詞，{ 名詞 } は名詞と解析された単語である．係り受け解析を行うことで,「赤い 彼の 髪」のように間に他の語が入っても特徴を抽出できるようにした．係り受け解析には Cabocha[3]を使用する．

ただし，抽出規則に合致した単語の品詞が連用形の場合は除外する.「静かに 手を動かす」という文の下線部では，連用形の「静かに」は，身体的特徴を表すのではなく，動作を修飾しているからである．

3.3.6 性格

人物の性格を抽出するために「基本的な性格表現用語のリスト」[10] を参考にして性格リストを作成した．性格リスト中の語と表層一致した文字列を人物の性格とする．ただし，係り受け解析の結果，動詞やサ変接続の名詞に係っている単語は除外する．例えば「ホームズは 静か に入り口の戸を閉める」という文の下線部では，動詞の「閉める」に係っている「静か」は，性格を表現しているのではなく動作を修飾しているからである．

3.4 人物情報のリスト作成

小説テキストから抽出した人名とその人名に付与された属性値を一覧できるように，表形式でまとめる．

4 人物情報の体系化

4.1 関連度の計算

本研究では，同じ場面に登場している人物の間には何らかの関係があると判定する．

「場面」とは，小説のストーリーに基づいて分割したテキストの断片である．しかし，場面を自動的に分割することは困難であるため，本研究では,「登場人物の入れ替わり」,「場所の変化」,「時間の経過」を場面変化の指標として人手で小説テキストを分割している．なお，分割の単位は 3.1 節で定義した文の単位で行う．すなわち，文の途中で場面は変わらないことを仮定している．

次に，テキスト中に人名が出現していることを存在の根拠とする．しかし，人名の出現だけで判定を行うと 2 つの問題が生じる．具体的には，人名が書かれていてもその場面にいるとは限らないことと，同一人物に対して異なる表記が使われることである．

[2] http://ygdrsl.s14.xrea.com/

[3] http://chasen.org/~taku/software/cabocha/

そこで，存在の確からしさとして「台詞情報」と「入退場情報」を利用する.「台詞情報」とは，会話文とその発言者の組である．発言している人物は場面にいる確からしさは高くなる.「入退場情報」とは,「来た」と「帰った」のように入退場が判る表現とその動作主の組である．

入退場表現は『角川類語新辞典 [7]』の「往来」項目を参考に収集した名詞と動詞のリスト（合計 87 語）との表層一致によって抽出する．退場表現があれば，動作主に対する存在の確からしさを低くする．入場表現があれば，動作主に対する存在の確からしさを高くする．なお,「帰って 来た」のように退場表現と入場表現を両方含む表現では下線部を引いた後半の単語「来た」に基づいて入場と判定する．

ある場面における存在の確からしさは，人名が出現した場合に 0.5 とする．人名が出現していなければ 0 とする．さらに，台詞があれば +0.3，入場表現があれば +0.2，退場表現があれば −0.2 する．例えば,「人名が出現して台詞のあった人物」の確からしさは $0.5+0.3=0.8$ である．さらに，入場表現があれば，+0.2 されて最大の 1 となる．ある人物に対する存在の確からしさをテキスト全場面分で総和した数値をその人物が登場した場面数とする．

また，人名の異表記に対処するために同一人物判定を行う．同一人物判定には二種類の処理がある．

一つ目は,「シャーロック・ホームズ」と「ホームズ」のように「姓名」と「姓と名のどちらか一方」の表現を一致させるための処理である．人名に含まれる「・」または「＝」の前後にある語を「姓」と「名前」とする.「姓」や「名前」と小説テキスト内にある他の人名が表層一致した場合は同一人物と判定する．ただし,「シャーロック・ホームズ」と「マイクロフト・ホームズ」のように同一の姓または名前を持つ人物が複数いる場合は，この処理を行わない．

二つ目は,「人名」と「人物を示す名詞」の表現に対処するための処理である．小説テキストでは，人名の末尾に付属した名詞が人名の代わりに使われることがある．例えば，以下のテキストでは，下線を引いた「グレゴリ」と「警部」は同一人物である．

> グレゴリ 警部は頷いた．ホームズは訊ねた.（略）
> 「はて」と，警部 は苦い顔をした.

そこで，対象の小説テキストから辞書を事前に作成して同一人物判定に使用する．まず，3.2 節で抽出した人名の直後に接続している名詞を全て収集する．例えば,「グレゴリ 警部 とロス 大佐 はホームズを待っていた」という文では，下線を引いた「警部」と「大佐」が収集される．ただし，収集した語のうち，複数の人名に対して出現している語は，曖昧性が生じるため削除する．例えば,「レストレード警部とグレゴリ警部がロス大佐を迎えに来た」という文では,「大佐」と「ロス」を対にして辞書に登録する．しかし,「警部」は複数の人物に対して使われているので削除する．

最後に，共起頻度が高い人物同士に高い関連度を与える．しかし，登場回数が多い人物同士は偶然共起している可能性があるため，Dice 係数を用いて人物の関連度を計算する．具体的には，人物 A と人物 B の関連度を式 (1) で計算を行う．

$$\frac{2\times(\text{人物 A と人物 B が共起する場面数})}{(\text{人物 A が登場する場面数})+(\text{人物 B が登場する場面数})} \quad (1)$$

4.2 人物相関図の描画

本研究ではノードを人物とし，人物関係をエッジとしたグラフによって人物相関図を表現する．関連度が高いほどノード間のエッジを短くする．グラフ作成には Touchgraph[4] を使用する．

5 評価実験

青空文庫[5]に収録されている作品から，英米文学の推理小説 4 件を対象に評価実験を行った．対象とした小説を表 1 に示す．

表 1: 評価実験に使用した小説

タイトル	黄色な顔	空き家の冒険
著者	コナン・ドイル	
文数	691	587
場面数	17	10

タイトル	白銀の失踪	モルグ街の殺人
著者	コナン・ドイル	エドガー・アラン・ポー
文数	738	838
場面数	19	10

人物情報抽出は，人手で判定した正解と本手法の結果を比較して精度と再現率で評価した．しかし，人物相関図の質を定量的に評価することは困難であるため，作成された人物相関図に対して目視で考察を行った．

5.1 人物情報抽出の評価

人物情報抽出の評価では，人名抽出と属性抽出についてそれぞれ評価を行った．表 2 に結果を示す．

人名抽出では，人物に具体名がなかったことと，形態素解析誤りのために抽出漏れが起こった．また，形態素解析誤りによって抽出誤りが起こった．

[4] http://www.touchgraph.com/
[5] http://www.aozora.gr.jp/

表 2: 人物情報抽出の精度と再現率 (%)

		黄色な顔	空き家の冒険
人名抽出	精度	35.3(6/17)	40.5(17/42)
	再現率	66.7(6/9)	73.9(17/23)
属性抽出	精度	10.0(1/10)	21.2(7/33)
	再現率	3.1(1/32)	13.5(7/52)
		白銀の失踪	モルグ街の殺人
人名抽出	精度	40.6(13/32)	53.3(16/30)
	再現率	72.2(13/18)	55.2(16/29)
属性抽出	精度	29.5(13/44)	33.3(3/9)
	再現率	28.9(13/45)	13.0(3/23)

属性抽出では,「ホームズ:探偵，男性」や「ワトソン:医師・看護師」などが正しく抽出された.

属性抽出における抽出漏れの原因は三つあった．以下,例と共に示す.

(1) 特徴を表す表現の周辺に人名が登場しない

身体的特徴： 青色の瞳，背の高い，色の白い

(2) 抽出規則の辞書に対応する語句が含まれていない

年代： 年輩

職業： 馬丁，調教師，靴直し，銀細工業，葬儀屋

性格： 臆病，狡猾

(3) 抽出規則で対応できない

年齢： 三十をちょっとすぎたぐらい

身体的特徴： 小柄，顔色がひどく青ざめて

抽出誤りの例として，他人の特徴を誤って抽出した事例があった．また,「深い爪の痕があって」という表現から,「深い爪」を身体的特徴として抽出した誤りの例があった.

5.2 人物相関図に関する考察

図 5.2 は『白銀の失踪』に対して作成された人物相関図である．探偵，警官，容疑者，被害者，依頼者といったストーリー上で重要な役割を持つ人物がグラフ中央（実線の楕円で囲っている）に配置された．他方で，特定の場面だけに登場する人物は相関図外延部で小さなグループを形成した（図 5.2 では点線で囲っている）.

表 1 に示した他の 3 作品に対しても、探偵などのストーリー上で重要な役割を持つ重要人物は相関図の中央に現れ，特定場面だけに登場する人物は外延部にグループを形成した.

相関図における人物の配置を利用することで，重要人物の判断が可能である．また，重要人物の行動に着目することであらすじ生成への応用に発展する可能性がある.

ただし，本研究の相関図からは人物間の関係（親子関係，友好関係，敵対関係など）はわからない．関係のラベル付けを自動的に行うことは今後の課題である．人物間の関係がラベル付けされることによって,「探偵と警官の仲が悪い（敵対関係である）推理小説を読みたい」といった検索が可能になる.

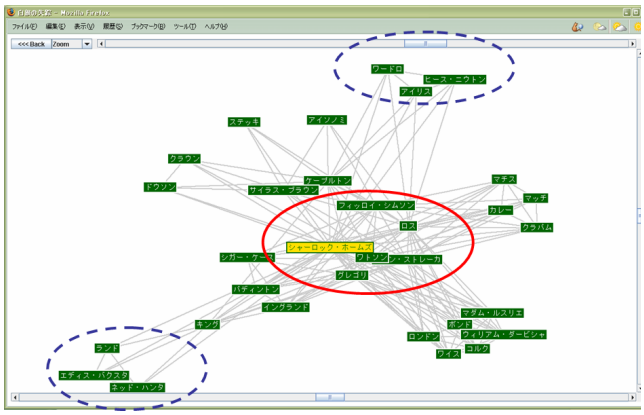

図 2: 『白銀の失踪』に対する人物相関図

さらに，あらすじ生成において,「ホームズは 友人 のワトソンと共に事件現場に向かった」という文における下線部のように，人物関係の描写を含めるような応用が可能になる.

6 おわりに

小説テキストは，ストーリーや登場人物に基づいて検索や分類されることがあるため，索引語に基づくモデル化だけでは限界がある．本研究は，人物と人物関係に着目して小説テキストをモデル化する手法を提案した.

抽出に使用する辞書や規則の大規模化と人物相関図における関係のラベル付けが今後の課題である.

参考文献

[1] 馬場こづえ, 藤井敦, 石川徹也. 小説テキストを対象としたジャンル推定と人物抽出. 言語処理学会第 11 回年次大会発表論文集, pp. 157–160, 2005.

[2] 馬場こづえ, 藤井敦, 石川徹也. 小説テキスト自動分類のためのジャンル推定と人物抽出. 第 4 回情報科学技術フォーラム講演論文集, pp. 67–70, 2005.

[3] 野崎広志, 中澤俊哉, 重永実. 物語理解におけるエピソード・ネットワークの構築. 情報処理学会論文誌, Vol. 30, No. 9, pp. 1103–1109, 1989.

[4] 中澤俊哉, 重永実. エピソードネットワークを用いた物語のあらすじ生成. 情報処理学会論文誌, Vol. 32, No. 10, pp. 1215–1224, 1991.

[5] 今誠一, 吉田文彦, 内田理, 菊池浩明, 中西祥八郎. 昔話の自動シナリオ化システムの構築. 言語処理学会 第 11 回 年次大会 論文集, pp. 317–320, 2005.

[6] 日外アソシエーツ（編）. 8 万人西洋人名よみ方綴り方辞典. 日外アソシエーツ, 1994.

[7] 大野晋, 浜西正人（編）. 角川類語新辞典. 角川書店, 1981.

[8] 中村明（編）. 人物表現辞典. 筑摩書房, 1997.

[9] 中村明（編）. 感覚表現辞典. 東京堂出版, 1995.

[10] 村上宣寛. 基本的な性格表現の収集. 性格心理学研究, Vol. 11, No. 1, pp. 35–49, 2002.