推理小説の難易度評価のための犯人推定

山本 聖也　奥村 紀之
香川高等専門学校　情報工学科

pornofan_tetra7717@yahoo.co.jp , okumura@di.kagawa-nct.ac.jp

1　はじめに

近年，著作権の切れた書籍が電子化され，誰でも簡単にそれらを Web 上で閲覧することができるようになった．それにより，推理小説の読者数も増加していると考えられる．しかし，推理小説には多くのジャンルが存在し，その推理の難易度も様々である．推理小説の難易度は犯人の予測しやすさによって決定される．

そこで本研究では，難易度評価のための犯人抽出を行う．分類器を構築することで，犯人が抽出されやすい小説や，抽出されにくい小説の傾向から，推理小説の難易度評価に繋げられるかを評価する．

2　関連研究

推理小説とは，推理作家江戸川乱歩によって「主として犯罪に関する難解な秘密が，論理的に徐々に解かれていく経路の面白さを主眼とする文学である [1]」と定義されている．基本的に，推理小説で起こる事件は殺人事件である．

推理小説では，物語の流れを追い，登場人物の行動や台詞などに基づいて犯人を推定する．西島らの研究 [2] では，文単位での繋がりを調べるために物語を結末から追跡し，命題間の論理的関係を抽出する手法を用いている．

また，台詞は犯人推定のための重要な要因であるため，その発話者を特定することが重要であると考えられる．Hua らの研究 [3] では，台詞の発話者を特定するために，抽出した人物名を元に，台詞内の呼格や発話者の交代パターンなどから，発話者と考えられる人物をランキング形式でモデル化する手法を用いている．

本研究では，小説に登場する人物の出現頻度に着目し，分類器を構築することで犯人抽出を行う．

3　人物名の解析方法

本研究では，登場人物の出現回数と物語の進度におけるその推移を考慮して犯人を推定する．適切に出現回数を抽出するためには，照応解析による代名詞の変換が必要となる．本節では，構文解析器 KNP を用いた代名詞変換について述べる．

3.1　代名詞等の変換

登場人物の出現回数を正確に求めるためには，作中に登場する代名詞や，役職名を対応する人物へ変換する必要がある．そこで，構文解析器 KNP の照応解析を用いて機械的に対応付けを行う．KNP は，与えられたテキストを形態素解析器 JUMAN で解析し，その解析結果を受けて照応解析を行っている．また，KNP には以下のオプションを付与して解析する．[4]

表 1: KNP の付与オプション

オプション	内容
anaphora	ルールに基づく共参照解析
	述語項構造解析
ne	CRF に基づく固有表現解析

KNP の解析結果より，小説内の代名詞や役職名を対応する登場人物名に変換する．以下の図 1，図 2 は「麻雀殺人事件 [1]」について，KNP と手動で代名詞変換を行い，各人物の出現割合を章分けに沿って調べた結果である．

この 2 つの結果より，KNP は手動と比較しても遜色ない代名詞変換が可能であると考えられる．この他に数編の小説で同様の実験を行ったが，同様の結果が得られているため，KNP による自動代名詞変換が可能であることが分かる．[5]

[1] http://www.aozora.gr.jp/cards/000160/card1220.html

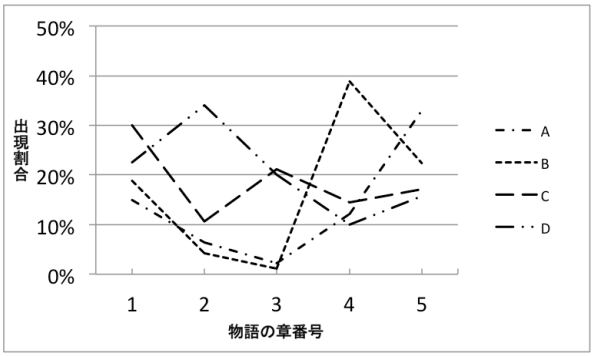

図 1: KNP を用いて代名詞変換した結果

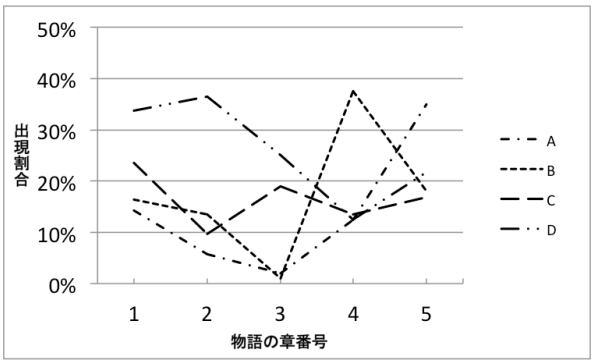

図 2: 手動で代名詞変換した結果

3.2 人名抽出方法

登場人物名の抽出には，JUMAN-7.0 と KNP を用いる．3.1 節の手法を用いて照応解析を行った小説テキストを入力として JUMAN で解析し，表 2 のパターンの出力結果が得られた単語を人名とする．また，JUMAN の既存の辞書に登録されている人名数だけでは不足していると考えたため，松茸用人名辞書 [6] をユーザ辞書として登録した．これにより，人名が 8 万語増加した．

表 2: 人名として抽出する JUMAN 出力パターン

フィールド	内容
品詞	名詞
品詞細分類	人名

人名の抽出について形態素解析のみで対応可能であるかを検証するため，実験的に JUMAN 単体の場合と，JUMAN と KNP を組み合わせた場合で，4 編の小説から人名の抽出を行った．その結果を表 3 に示す．JUMAN 単体で人名を抽出した場合は平均で 56 %程度しか抽出できなかったが，KNP と組み合わせた場合は，平均 97 %以上の人名抽出が可能だった．この結果より，取りこぼしなく人名を抽出するためには構文情報も用いる必要があると分かった．

表 3: 各手法の人名抽出精度

作品名	JUMAN	JUMAN,KNP
愚人の毒	0.727	1.000
カンカン虫殺人事件	0.636	0.909
謎の咬傷	0.546	1.000
麻雀殺人事件	0.333	1.000

4 犯人の出現傾向

青空文庫 [2] から推理小説を 100 編収集し，犯人の出現傾向に関する調査を行った．

まず，収集した推理小説を 3.1 節，3.2 節の手法で解析し，登場人物数を調査した．その結果が表 4 である．なお，ここでの登場人物数とは，3.2 節の手法により人名と判断された単語タイプ数である．

表 4: 小説の登場人物数

最少	9 人
最多	357 人
平均	41.29 人

次に各登場人物の物語全体での出現回数を調査した．その結果を元に，犯人が物語全体での総出現回数で 1〜5 位の各位にいる確率を求めた．それらを以下の表 5 に示す．

表 5: 総出現回数の各位に犯人が属する確率

順位	犯人
1 位	16.35 %
2 位	26.92 %
3 位	13.46 %
4 位	16.35 %
5 位	4.81 %
6 位以下	22.11 %

この結果を見ると，犯人は総出現回数において，全登場人物の内第 2 位に属する確率が最も高く，27 %という結果になった．

さらに，各小説を文数で 10 分割し，それぞれの部分における各登場人物の進度出現割合と，個人出現割

[2] http://www.aozora.gr.jp/index.html

合を調べた．これにより，犯人が物語のどのタイミングで頻繁に出現し，どのタイミングで出現頻度が減少するか，などの細かい出現情報に関する特徴が得られると考えた．

ここで進度出現割合とは，ある章・進度における全登場人物の総出現回数を 100 %とした時の，各登場人物の占める割合と定義し，個人出現割合とは，物語全体におけるある人物の総出現回数を 100 %とした時の，ある進度での出現割合と定義する．

全小説の犯人の，各部分での進度出現割合と個人出現割合の平均を以下の表 6 に示す．分割部分 1 とは，物語を文数で分割した際の物語冒頭から，全文数の 10 %を経過するまでの範囲を意味する．

表 6: 犯人の平均進度出現割合と平均個人出現割合

分割部分	進度出現割合	個人出現割合
1	7.89 %	5.06 %
2	11.07 %	6.00 %
3	10.25 %	6.08 %
4	10.43 %	6.70 %
5	9.27 %	5.57 %
6	10.69 %	7.32 %
7	9.36 %	6.40 %
8	10.26 %	8.08 %
9	13.56 %	11.10 %
10	18.57 %	21.47 %

この結果を見ると，犯人は物語終盤に出現割合が大きく上昇することが分かる．この原因として，推理小説では，探偵役が他の登場人物に向かって自分の推理を発表し，犯人を暴くシーンが物語終盤に置かれることが多いことが考えられる．このシーンで探偵は犯人の行動や動機などを詳しく説明することが多く，その際主語に犯人の名前が使われる．よって，推理小説では物語終盤に犯人の出現割合が上昇する傾向にあると考えられる．

5 分類器の構築

4 節の調査結果より，登場人物の出現情報を素性としてベクトルを生成し，犯人を自動抽出するための分類器を構築する．SVM は libsvm-3.20 を用い，SVM タイプは C-SVM，カーネル関数は多項式カーネルを用いる．

5.1 学習データ

分類器の学習・評価データとして用いるのは，青空文庫から収集した 100 編の推理小説の登場人物の出現情報である．犯人に該当する人物のベクトルを正例，それ以外の人物のベクトルを負例としている．また，複数の犯人が出現する作品に関してもすべて犯人は正例としてベクトルを作成している．用意したデータの正例と負例の数は以下の表 7 の通りである．

表 7: 分類器に用いる事例数内訳

正例	負例
104	4026

5.2 交差検定による実験

1 編の小説の評価データに対し 99 編の小説の学習データを用いて，以下の 7 パターンのベクトルによる交差検定を行い，比較した．その結果として得られた犯人分類器の正答率，再現率，適合率をそれぞれ表 8 に示す．正答率は式 (1)，再現率は式 (2)，適合率は式 (3) で表す．

1. 物語を 10 分割した各部分における個人出現割合と，物語全体での出現割合の 11 次元ベクトル
2. 物語を 10 分割した各部分における進度出現割合と，物語全体での出現割合の 11 次元ベクトル
3. 物語を 10 分割した各部分における個人出現割合と進度出現割合，物語全体での出現割合の 21 次元ベクトル
4. 物語を 15 分割した各部分における個人出現割合と進度出現割合，物語全体での出現割合の 31 次元ベクトル
5. 物語を 20 分割した各部分における個人出現割合と進度出現割合，物語全体での出現割合の 41 次元ベクトル
6. 物語を 20 分割した各部分における個人出現割合と進度出現割合，物語全体での出現割合の 51 次元ベクトル
7. 物語を 30 分割した各部分における個人出現割合と進度出現割合，物語全体での出現割合の 61 次元ベクトル

$$\text{正答率} = \frac{\text{正しく分類できた事例数}}{\text{全事例数}} \quad (1)$$

$$\text{再現率} = \frac{\text{正しく犯人と分類できた事例数}}{\text{全正例数}} \quad (2)$$

$$\text{適合率} = \frac{\text{正しく犯人と分類できた事例数}}{\text{犯人と分類した事例数}} \quad (3)$$

最小分割数を 10 としたのは，予備実験として物語を 2〜10 分割して，パターン 3 のベクトルを生成し，各分割数について 10-分割交差検定を行った結果，最も正答率が高かったためである．更に分割数を増やした場合の結果を調べるため，15 分割から 30 分割まで 5 分割区切りで同様の交差検定を行った．

表 8: 小説毎の交差検定結果

パターン	正答率	再現率	適合率	F 値
1	85.38 %	41.00 %	7.61 %	0.128
2	94.05 %	26.00 %	20.47 %	0.229
3	92.59 %	27.00 %	13.85 %	0.183
4	93.10 %	27.00 %	16.77 %	0.207
5	93.09 %	36.00 %	21.05 %	0.266
6	93.21 %	26.00 %	17.11 %	0.206
7	94.01 %	29.00 %	22.48 %	0.253

この結果を見ると，個人出現割合のみを用いたパターン 1 の結果が最も高い再現率を記録していることが分かる．しかし，正答率と適合率に関しては他のパターンよりも低くなっている．パターン 2 は，パターン 1 とは逆に正答率と適合率が最も高くなっているが，再現率は低下している．パターン 1 と 2 を組み合わせたパターン 3 では，全ての値についてパターン 1 と 2 の中間になった．適合率はパターン 7 が最大だが，パターン 5 もあまり変わらない値を出し，F 値を求めた結果パターン 5 が最も高い値を出した．パターン 4 と 6 は，正答率は他のパターンと変わらないが，再現率・適合率両方について減少するという結果になった．この結果から，犯人分類器のベクトルにはパターン 5 のベクトルを用いるのがふさわしいと考えた．

6 まとめ

本稿では，小説テキストを機械的に解析し，そこから得られる登場人物の出現情報をもとに犯人を推定する方法について述べた．

本研究では，登場人物の個人出現割合と進度出現割合を元にベクトルを用いて，表 8 に示したとおりの分類器を作成することができた．今後は，小説テキストから得られる他の特徴量についても調査し，分類器の素性について深く吟味することで，分類器の精度を高めていく．

本稿で構築した分類器は青空文庫に掲載されている著作権切れの古い小説である．そのため，構築した分類器が最新の小説にも適用可能であるかの検証も進めていく必要がある．

参考文献

[1] 島崎博. 幻影城. 絃映社.1951.

[2] 西島恵介, 神山文子, 藤田米春.(1997). 短編推理小説の論理構造の分析と推理. 情報処理学会全国大会講演論文集. 第 55 回平成 9 年後期 (2).pp57-58.

[3] H.Hua, D.Barbosa and G.Kondrak, "Identification of Speakers in Novels," in The 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp1312-1320.

[4] 笹野遼平, 河原大輔, 黒橋禎夫, 奥村学.(2013). 構文・述語項構造解析システム KNP の解析の流れと特徴. 言語処理学会第 19 回年次大会発表論文集.pp110-113.

[5] S.Hiraoka and N.Okumura.The Difficulty Estimation Method for Mystery Novels using Context Analysis.SCIS-ISIS2014.PP.1296-1301.

[6] 松茸用人名辞書.http://www.vector.co.jp/vpack/browse/person/an006131.html, 1998, (2014 年 5 月 20 日アクセス).

[7] 馬場こづえ, 藤井敦.(2007). 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.pp574-577.

[8] 相良直樹, 砂山渡, 谷内田正彦.(2004). 重要文抽出を利用したテキストからのストーリー抽出. 情報処理学会研究報告.2004-NL-164.pp159-164.