小説テキストを対象としたジャンル推定と人物抽出

馬場こづえ † 藤井 敦 †† 石川徹也 ††
† 筑波大学図書館情報専門学群
†† 筑波大学大学院図書館情報メディア研究科
{m218,fujii,ishikawa}@slis.tsukuba.ac.jp

1 はじめに

インターネットの普及によって大量のテキスト情報が氾濫している．これらを分類し，秩序を与えることで利便性を高めることができる．しかし，人手による分類は，時間的，経済的コストがかかる上，客観性や一貫性に欠ける．そこで，テキスト自動分類が研究されている．

テキストには，新聞，論文，特許などの「情報伝達テキスト」や，小説，日記，エッセイなどの「娯楽・芸術テキスト」がある．

テキスト自動分類では情報伝達テキストを対象とした研究が多い．娯楽・芸術テキストを対象とした研究には，著者推定 [1, 2]，ストーリー抽出 [3]，物語理解 [4] があるものの，自動分類に関する研究事例は少ない．

近年，著作権が切れた小説の電子化が行われている．電子出版も普及してきた．また，アマチュアが書いた小説が Web 上で公開されており，それらを対象とした検索サイトが複数存在する．例えば,「楽園[1]」ではジャンルによる検索,「世界樹の下[2]」では登場人物の特徴による検索が可能である．

以上より，Web 上の小説テキストを自動分類する需要が増えている．小説には,「著者」,「ジャンル」,「世界観」,「読者層」のように分類の観点が複数存在する．また,「ストーリー」や「登場人物」のように情報伝達テキストでは通常考慮しない観点がある．

本研究は,「ジャンル」と「登場人物」による小説の分類を目的とし，小説テキストに対するジャンル推定と人物抽出の手法を提案する．

2 ジャンル推定手法

2.1 概要

「ジャンル」とは，書店で書籍を配架するような分類であり,「SF」や「ファンタジー」などがある．ジャンルによる分類は，情報伝達テキストの分類と技術的な共通点が多い．そこで既存の分類手法を応用する．

近年，サポートベクトルマシン（SVM）[5] がテキスト分類に応用されている [6, 7]．SVM は特徴量ベクトルの次元数が増えても計算量が増えないため，大量の特徴量を扱う目的に適している．

[1]http://novel.pekori.to/
[2]http://ygdrsl.s14.xrea.com/

情報伝達テキストは，内容を簡潔に伝えるため，本文は短く，使用する語は統制される傾向がある．

それに対して，小説テキストは長さに関する制約が緩やかで，また，一つの概念を様々な表現で言い換える．すなわち，小説中の語を特徴量として扱う場合，情報伝達テキストよりも次元数が増える可能性がある．そこで，SVM は小説テキストの分類に有効であると考えた．

小説テキスト中の語を特徴量として使用するために形態素解析を行う．また，TF.IDF 法で特徴量の語に重み付けを行う．テキスト d における語 t の出現頻度を $tf(t,d)$ とし，全テキストデータの中で t を含むテキスト数を $df(t)$，テキストデータの総数を N とすると，d における t の重みを式 (1) で計算する．

$$(\log tf(t,d)+1)\cdot(\log\frac{N}{df(t)}+1) \qquad (1)$$

特徴量抽出に用いるテキスト集合は，一定の質が保たれ，大量にあることが好ましい．そこで「BOOK」データベース 1999 年版 [8] を学習データとして利用する．当データベースには，51171 件の書籍について，書名，著者，出版社，書店的分類，要旨，目次，などの情報が含まれている.「書店的分類」をジャンルとし,「要旨」のテキストを形態素解析して特徴量となる語を抽出する．

2.2 多値分類への拡張

SVM は二値分類であるため，ジャンル推定のために多値分類へ拡張する．多値分類への拡張方法には one-versus-rest 法と pairwise 法がある [9].

one-versus-rest 法はカテゴリが N 個あるとき，あるカテゴリに該当するか，否かという分類器を N 個作成し，それぞれの分類器で分類を行う．SVM は分類の確かさを示すスコアを計算することができるので，N 回の分類の中でスコアが最も高かったカテゴリを選択する．

pairwise 法はカテゴリが N 個あるとき，${}_NC_2$ 個の分類器を作成し,「リーグ戦」によってカテゴリを決定する．例えば，A，B，C の三つのカテゴリがあったとき，A 対 B，A 対 C，A 対 B という分類器を作成する．それぞれの分類器で分類を行い，各カテゴリに対するスコアを計算して，スコアの合計が最も高かったカテゴリを選択する．

3 人物抽出手法

3.1 概要

登場人物で小説を分類するためには，本文から登場人物とその特徴を抽出しなければならない．また，ストーリー抽出や物語理解において，登場人物は重要である．

本研究で提案する人物抽出手法の概要を以下に示す．

(1) 小説テキストを形式段落で分割する．

(2) 各段落から人名を抽出する．

(3) 人名の周辺文脈から特徴を表す表現を抽出する．

「エリックの 金の髪 が宙を舞う．彼 は器用に着地すると 少女 の手をとって，一目散に逃げ出した.」のように，下線部分を「エリック」という人物の特徴として抽出する．

(4) (2) と (3) で抽出した人名とその特徴を人物情報としてまとめる．

以下，3.2 節〜3.4 節で上記の手順 (1)〜(3) についてそれぞれ説明する．

3.2 小説テキストの分割

人名の周辺文脈から特徴表現を抽出するため，周辺文脈の範囲を制限する必要がある．本研究では形式段落を文脈の範囲とした．

しかし，文章の区切り方は著者によって様々である．本研究で使用する小説テキストは Web から収集する．そこで，段落の判断に HTML タグを使用する．段落を表す ， と，改行を表す
でテキストを分割する．HTML 形式ではないプレーンテキストの場合は，改行記号でテキストを分割する．

3.3 人名抽出

人名辞書を利用し，辞書に載っている人名を小説テキストから抽出する．人名辞書は「怪しい人名辞典[3]」を利用した．この辞典では，英語，ドイツ語，フランス語，イタリア語，ロシア語の個人名（first name）が 2780 件が集められている．

人名辞書に載っていない語は形態素解析によって抽出する．人名として抽出された語のテキスト全体における出現回数を f，小説テキストに含まれる形式段落数を L とした場合，$\frac{f}{L}$ が閾値よりも大きい場合にその語を人名と判定する．

3.4 特徴表現の抽出

3.3 節の手法で抽出された人名の周辺文脈から，人手で作成した規則に表層一致した表現を抽出し，該当するカテゴリを特徴として選択する．対象とする特徴は「性別」,「年齢」,「年代」,「職業」,「身体的特徴」,「性格」である．ただし，周辺文脈は人名が現れる段落内に限定する．

[3]http://www5d.biglobe.ne.jp/˜ros/aya.htm

性別　性別がわかる語（男性，母，叔父など）や性別固有の一人称（俺，わしなど）が含まれていれば，該当する性別を選択する．

ただし，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性別を選択する．

年齢　「17 歳」や「三十五才」といった表記を規則によって抽出し，そのまま特徴とする．数字表記の次に「才」または「歳」の字が現れた場合を年齢表記とする．数字表記は「1」のような半角アラビア数字,「１」のような全角アラビア数字,「一」,「百」といった漢数字とする．

年代　人間の一生を「乳幼児期」,「少年期」,「青年期」,「中年期」,「老年期」のカテゴリに区分する．各年代を示す語をリスト化し，リスト中の語に表層一致した場合，該当する年代を選択する．また,「20 代」といった年代があれば，そのまま特徴とする．

一人の人物に対して一つの年代のみを選択する．複数の年代が出現した場合はテキスト全体で最高頻度のカテゴリを選択する．ただし，年齢と年代の対応は絶対的な基準がないため，年齢から年代を推測することはしない．

職業　「世界樹の下」を参考に「剣士・騎士・戦士」といった職業リストとその特徴語を作成した．特徴後と表層一致した場合，該当する職業を選択する．

身体的特徴　髪や瞳の色，声，体格など，容姿に関する特徴を抽出する．カテゴリ数に制限はなく，抽出した特徴をそのまま人物に付与する．

抽出規則は三つあり，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」は既存の辞書 [10, 11] を参考にして作成した，63 語からなるリストである．{ 形容詞 } は形態素で形容詞と解析された語である．{ 名詞 } は形態素解析で名詞と解析された語である．

この規則によって,「髪は黒い」,「明るい声」,「緑の瞳」といった特徴が抽出される．

性格　性格表現リスト中の語に表層一致した文字列を人物の性格とする．性格表現リストは既存のリスト [12] を元に作成した．このリストは広辞苑から収集した語と性格表現に関する語のリストから 934 語を収集したものである．本研究では 934 語の中から名詞，形容詞の一部を用いた．また,「真面目」と「まじめ」のような異表記語を追加し，最終的に 208 語からなるリストを作成した．

4 評価実験

4.1 ジャンル推定手法の評価

4.1.1 方法

「楽園」2001 年登録分の小説テキストをテストデータとして利用した．リンク切れや，未完結の小説は人手で

除き，758 件を収集した．学習データは「BOOK」データベースから作成した．

しかし，楽園と「BOOK」データベースではジャンル体系に差異がある．また,「BOOK」データベースには娯楽・芸術テキスト以外に実用書や統計書など含まれている．そこで，人手でジャンルの対応付けを行い，対象ジャンルが該当しないデータは除外した．最終的な 6 ジャンルと該当するデータ件数を表 1 に示す．

表 1: 楽園と「BOOK」データベースのジャンル内訳

ジャンル	楽園	「BOOK」DB
童話	22	613
現代物	121	970
恋愛小説	154	314
ミステリー	35	1558
歴史小説	10	605
SF・ファンタジー	227	677
合計	558	4737

ジャンル推定は分類の正解率によって評価する．正解率は，正しいジャンルに分類されたテキスト数をそのジャンルに該当するテキスト数で割って計算する．

品詞選択と特徴量への重み付けによる正解率の影響を評価するため，以下に示す 4 つの手法を比較した．ここで，内容語とは，動詞，名詞，副詞，形容詞，未知語である．

(a) 全ての語を特徴量とし，重みは全て 1 とする．

(b) 内容語を特徴量とし，重みは全て 1 とする．

(c) 全ての語を特徴量とし，TF.IDF で重み付けをする．

(d) 内容語を特徴量とし，TF.IDF で重み付けをする．

学習と分類には TinySVM[4]を使用し，形態素解析には Chasen[5]を使用した．多値分類への拡張法として one-versus-rest 法と pairwise 法があるものの，予備実験で正解率が高かった one-versus-rest 法の結果だけを示す．

4.1.2 結果と考察

実験結果を表 2 に示す．表中の (a)〜(d) は，4.1.1 節で説明した手法 (a)〜(d) の正解率である.「平均」はまずジャンルごとの正解率を計算し，それらを平均した「マクロ平均」である．そこで，特定のジャンルに対する正解率だけが高くても，平均は高くならない．

(a) と (b) を比較すると,「童話」,「歴史小説」,「SF・ファンタジー」では特徴量の品詞を限定した (b) の正解率が高く，他のジャンルでは全形態素を特徴量とした (a) の正解率が高かった．よって，ジャンルによって有効な特徴量が異なることがわかった．

「ミステリー」における (b) と (d) 以外では，TF.IDF による重み付けのために正解率が下がった．重み付けが効果的でなかった原因として，学習データ（要旨）とテストデータ（小説テキスト）で文章の書き方や語の頻度分布が異なることが考えられる．

[4]http://chasen.org/~taku/software/TinySVM/

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

表 2: ジャンル推定の正解率 (%)

	(a)	(b)	(c)	(d)
童話	86.4	95.5	22.7	0
現代物	17.4	12.4	4.1	71.1
恋愛小説	22.7	11.0	93.5	30.5
ミステリー	34.3	14.1	0	2.9
歴史小説	70	100	0	0
SF・ファンタジー	16.5	71.4	7.9	9.7
平均	40.0	45.4	21.4	19.0

この点について検討するために，楽園データを二等分して，一方を学習データ，もう一方をテストデータとして実験した．手法 (a)〜(d) の正解率を表 3 に示す．

「BOOK/楽園」は「BOOK」データベースで学習し，楽園データでテストした正解率である．

「BOOK/BOOK」は「BOOK」データベースを二等分して学習用データ，テストデータとした正解率である．「BOOK/BOOK」では，(a)〜(d) 全てで正解率が向上し，SVM による新聞記事分類 [6] とほぼ同等であった．「BOOK」データベースのテキストは要旨である.「楽園」の小説テキストと比べると文字数は少なく，語も統制されているため，分類精度が高くなったと考えられる．

「BOOK/楽園」と「楽園/楽園」を比較すると,「楽園/楽園」では，重み付けを行った (c), (d) の正解率が上がった．よって「BOOK/楽園」において (c)，(d) の正解率が低い理由は学習データとテストデータにおける語の頻度分布の違いである．

また,「楽園/楽園」の正解率が低い原因として，データ件数が少ないことが考えられる．今後は実験の規模を拡張し，この点について調査する必要がある．

表 3: 学習・テストデータの違いによる正解率 (%) の比較

	(a)	(b)	(c)	(d)
楽園/楽園	27.4	27.2	31.5	28.5
BOOK/楽園	40.0	45.4	21.4	19.0
BOOK/BOOK	77.7	75.0	77.0	73.7

4.2 人物抽出手法の評価

4.2.1 方法

人物抽出では以下の二点を評価した．

- 登場人物の人名が正しく抽出されたか
- 登場人物の特徴が正しく抽出されたか

「世界中の下」から収集した小説 5 件を対象に評価した.「世界樹の下」では，著者が登録した登場人物の名前，年齢，職業，容姿，性格などによって小説テキストが分類されている．

4.2.2 結果と考察

表 4 に人名抽出の実験結果を示す.「登録」は著者が登録した人名に対する抽出精度である．括弧内は（著者が登録した人物で抽出された人物数/登録人数）である．

「登録＋追加」は著者が登録しなかった登場人物について第三者が追加判定した場合の結果である．ここでは，「名前が出現している人物」と「台詞が一回以上ある人物」を人物と判定した.「抽出誤り」は，人名ではないものが抽出された件数である．

著者が登録した人物に対する抽出率は平均 47.6% 追加判定も含めた場合の抽出率は平均 15.5%だった．

抽出漏れの原因は，人名辞書の規模，形態素解析誤り，出現頻度，人名が出現しないの四点だった．抽出誤りの原因は形態素解析誤りだった．

表 4: 人名の抽出率

	小説 A	小説 B	小説 C	小説 D	小説 E
文字数	38617	57679	27659	78376	235151
登録	0% (0/8)	50% (1/2)	100% (2/2)	100% (6/6)	33.3% (1/3)
登録 +追加	13.3% (2/15)	13.3% (1/13)	27.3% (3/11)	21.1% (8/38)	9.1% (3/33)
抽出誤り	0	3	1	1	1

小説 A〜E における人物特徴抽出の結果を表 5 に示す．表 4 で人名の抽出漏れが多かったため,「世界樹の下」に登録されている人名を人名辞書に加えて実験を行った．「世界樹の下」には年代は登録されていないため,「0〜5 歳：乳幼児期」,「6〜15 歳：少年期」,「16〜40 歳：青年期」,「41〜60 歳：中年期」,「60 歳〜：老年期」として評価した．

「再現率」は登録された特徴が抽出された割合を示し，括弧内の数字は（登録された特徴が抽出された件数/「世界樹の下」に登録された特徴の件数）である．

「精度」は，抽出された特徴の正解率を示し，括弧内の数字は（正しく抽出された件数/抽出された特徴の件数）である．人物特徴抽出では，身体的特徴，性格に対する抽出誤りと抽出漏れが多かった．

表 5: 人物特徴抽出の再現率 (%) と精度 (%)

人物特徴	再現率 (%)	精度 (%)
性別	61.9(13/21)	61.9(13/21)
年齢	25.0(5/20)	62.5(5/8)
年代	–	5.3(1/19)
職業	25.0(8/32)	30.0(9/30)
身体的特徴	6.6(6/91)	12.0(22/183)
性格	3.4(4/117)	15.7(19/121)

4.2.3 具体例

身体的特徴抽出の成功例，抽出誤り例，抽出漏れの例を以下に示す．

- 成功：青い瞳，銀の髪，口調がつっけんどん
- 抽出誤り：相手の首，爪が容赦，薄い目
- 抽出漏れ：微笑みをたやさない

抽出誤りは,「相手の首に手を当てる」,「爪が容赦なく襲う」,「薄く目を開ける」など身体的特徴とは関係ない体言や用言を修飾している語を誤って抽出したことと，別人の特徴を誤って抽出したことが主な原因だった．

性格では「真面目」,「クール」といった特徴の抽出が正しくできた．しかし,「優しく話す」から「優しい」,「弱々しい声」から「弱々しい」というように用言を修飾している語や身体的特徴を表している語を誤って抽出した．また，別人の特徴を誤って抽出した．

抽出漏れの原因の一つは身体的特徴や性格の表現の周囲に人名が出現していなかったことである．

5 おわりに

小説テキストの多面的な分類を目的としてジャンル推定と登場人物の抽出手法を提案し，実験によって評価した．ジャンル推定では，ジャンルによって有効な特徴量が異なることと，学習データとテストデータの語の頻度分布の差異で分類精度が変化することを明らかにした．人物抽出では，主要な特徴を抽出することができたものの，抽出の誤りと漏れがあり，その傾向について分析した．

現在は，ジャンル推定と人物抽出は全く独立している．しかし，ジャンル固有の人物特徴が存在する可能性があるため，今後はジャンル推定と人物抽出の統合も必要である．また，ストーリー抽出などへ応用することも研究課題である．

参考文献

[1] 松浦司, 金田康正. n-gram の分布を利用した近代日本語文の著者推定. 計量国語学, Vol. 22, No. 6, pp. 225–238, 2000.

[2] 吉田篤弘, 斎藤博昭. サポートベクトルマシーンを用いた著者判別における有効素性推定. 言語処理学会 第 9 回年次大会発表論文集, pp. 513–516, 2003.

[3] 相良直樹, 砂山渡, 谷内田正彦. 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, pp. 159–164.

[4] 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.

[5] Cortes, C. and Vapnik, V. Support Vector Networks. *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[6] 平博順, 春野雅彦. Support Vector Machine によるテキスト分類における属性選択. 情報処理学会論文誌, Vol. 41, No. 4, pp. 1113–1123, 2000.

[7] 高村大也, 松本裕治. SVM を用いた文書分類と構成的機能学習法. 情報学会論文誌 データベース, Vol. 44, SIG 3（TOD 17）, pp. 1–9, 2003.

[8] 「BOOK」データベース　1999 年版. 日外アソシエーツ.

[9] 山田寛康, 松本裕治. Support Vector Machine の多値分類問題への適用法について. 情報処理学会研究報告, 2001-NL-146, pp. 33–38, 2001.

[10] 中村明（編）. 人物表現辞典. 筑摩書房, 1997.

[11] 中村明（編）. 感覚表現辞典. 東京堂出版, 1995.

[12] 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.