小説を対象とした文書分類手法の検討

三輪拓也　　松本忠博
岐阜大学 大学院 自然科学技術研究科
{miwa, tad}@.mat.info.gifu-u.ac.jp

1　はじめに

文書分類は，与えられた文書を指定したクラスのいずれかに分類する問題であり，分類器としてサポートベクターマシン（SVM）やニューラルネットワークを使用した研究が盛んに行われている．しかし，分類対象はニュース記事などの情報伝達を目的とした文書であることが多く，小説をジャンルによって自動分類する研究はまだ少ない．

近年では電子書籍や Web 小説などで小説を入手しやすくなり，西原ら [1] の登場人物の関係抽出や高田ら [2] のオンライン小説推薦手法など，小説を利用した研究が多く行われるようになった．

小説のジャンル推定の精度が向上すれば，読者は未知の小説に対して好みのジャンルかどうか判断ができ，Web 小説サイトの運営者は作品に対して客観的にジャンルを分類できるようになる．しかし，小説は一つの作品の中に，SF，ホラー，恋愛など複数の要素が含まれる場合が珍しくないなど，ニュース記事等と比較して分類が難しい面がある．

本研究では小説を対象とした文書分類を行った馬場ら [3, 4] の手法を基に，文章の前処理，特徴語の抽出方法，特徴語の選別などの改良を加え，Web 上から取得した小説を対象に評価実験を行い，分類精度の比較を行った．

2　分類対象としての小説の特徴

小説は，長さ（短編，長編），発表形式（連載，書き下ろし，オンライン），著者，内容（ジャンル，純文学／大衆文学），国（日本文学，英米文学）などによって分類されるが，本研究では一般的なジャンルによる自動分類を考える．

文書分類における従来の研究では，ニュース記事や論文などの情報伝達を目的とした文書を対象としたものが多い．これらの記事は読者に読みやすく，伝わりやすくする必要があるため，客観的な事実や意見を一定の語彙の範囲で簡潔に述べたものが多く，著者による表現の違いが出にくいため，著者に依存せず，文書中の語による分類（ジャンルの推定）が比較的行いやすいものと考えられる．

だが，芸術作品である小説では表現上の制約が少なく，語の選択，文体は自由であり，編集者等によって大きく修正されることも少ないと想像される．そのため小説には，ジャンルごとの特徴だけでなく，著者ごとの特徴が現れる．また，小説はジャンルが重複する作品が存在するため，従来使用されているデータよりも分類が難しいと予想できる．

3　先行研究

馬場ら [3] は，特徴量とする語の種類と重み付けの方法を変化させて，SVM（one-versus-rest 法）による小説の分類精度を比較した．語の種類については，すべての語を特徴量とした場合と，内容語（動詞，名詞，副詞，形容詞，未知語）のみを特徴量とした場合，重み付けについては，0 か 1 の 2 値とした場合と，tf-idf を用いた場合の合わせて 4 通りの方法で分類を試み，正解率を比較している．

図書内容情報データベース「BOOK」に含まれる小説の要旨を学習データ（4,737 件），ネット小説検索サイト「楽園」から収集した小説をテストデータ（558 件）として，「童話」「現代物」「恋愛小説」「ミステリー」「歴史小説」「SF・ファンタジー」の 6 ジャンルに分類する実験を行った結果，内容語のみを特徴量とし，重みを 0/1 とした場合に最も高い正解率（45.4%）を得た．学習データは要旨であり，ジャンルによらず文体はほぼ一定で，テストデータである小説の文章とは性質が異なる．

学習データとテストデータの両方に小説を用いた場

合と，両方に要約を用いた場合の分類も行っており，結果は，小説の場合は正解率 31.5%（すべての語と tf-idf）が最高であり，要旨の場合は 77.7%（すべての語と 2 値重み）が最高であった．

また馬場ら [4] は，小説のジャンル推定では，学習データとして，分類対象と同じ性質を持つ少量のデータ（小説）と，異なる性質を持つ大量のデータ（要旨）を併用することで，小説だけ，あるいは，要旨だけを学習データとした場合より高い正解率が得られることを示した．このとき，正解率の最大値は 58.2% で，すべての語を特徴量とし，重みを 0/1 としたときに得られている．

4　分類手法の改良

先行研究では学習データとして，大量の要約で少量の小説を補っていたが，本研究では要約は用いず，学習データ，テストデータともに小説を用いる．基本的に先行研究と同様，小説中の語を特徴量とし，重み付けに tf-idf などを用いて，SVM による分類を行うが，分類精度の向上を図るために，学習データ・テストデータとなる小説に対して以下に述べる処理を加える．

（手法 1）文書に対する前処理

文章をベクトル化する前に以下の処理を行う．

- 文字種を全角に統一する
- 数字を 0 に置きかえる

文字種が違うと同じ単語でも別の単語として認識してしまうため，全ての単語の文字種を全角に統一する．数字は全ての文章に共通して多く出現するため，分類の際に特徴語としてあまり役に立たないと判断した．ただし，漢数字の場合，「一緒」のように「漢数字 + 漢字」で構成する特徴語が存在するため，形態素解析の結果が「数」となる数字のみを置きかえる．

（手法 2）特徴語の抽出

小説では多くの固有名詞が出現する．だが，固有名詞は形態素解析用の辞書に存在せず，分割して解析されてしまうことが多いと予想できる．よって，連続して「名詞」と解析された単語を 1 つの単語として抽出する．

（手法 3）特徴語リストの作成

小説は基本的に文字数制限がないことや固有名詞が多く出現することから特徴語が多くなると予想できる．だが，出現頻度が低すぎる特徴語は学習には役に立たず，ただ学習時間が増加してしまう．また，逆に出現頻度が高い単語もあまり学習に役に立たない．よって，収集した全作品を対象に，各ジャンルの出現頻度が 30 %以上かつ全体の出現頻度が 80 %未満の単語を抽出した特徴語リストを作成し，リストにある単語のみを特徴語として使用する．

（手法 4）使用する作品の選別

小説には複数のジャンル要素を含んだ作品が多く存在する．複数のジャンル要素を含む小説が学習データに含まれていると，分類精度の低下につながることが予想される．そこで，複数のジャンルの要素を含む作品を学習・テストデータから除外する．

5　評価実験と考察

5.1　実験方法と評価

形態素解析は MeCab[*1]，形態素解析用の辞書は mecab-ipadic-NEologd[*2]，分類器は SVM を使用し，5 分割のクロスバリデーションで行った．先行研究の手法をベースラインとして，提案手法の結果を比較する．なお，先行研究における手法は著者が収集した作品で行った結果，すべて語と内容語を特徴語として比較したとき，差が 0.2 %と正解率にほぼ差がなかったため，内容語を特徴語とし，重みは tf-idf とした手法とする．提案手法は，先行研究の手法に手法 1〜3 の改良を行った提案手法 1，手法 1〜4 の改良を行った提案手法 2 とする．また，比較のため，先行研究の手法で新聞記事の分類も行う．

5.2　使用データ

小説のデータは小説を読もう[*3]から「推理」,「歴史」,「ホラー」,「SF」,「恋愛」,「ファンタジー」の 6 つのジャンルから合計 23189 作品，新聞記事のデータは 2005 年毎日新聞コーパスから「社説」,「国際」,「経済」,「家庭」,「読書」,「芸能」,「スポーツ」,「社会」の 8 つのジャンルから 8800 文を収集し，両データとも各ジャンル 1100 作品・文を実験で使用し，小説は文字数が 3〜10 万字の作品とする．また，提案手法 2 においては，使用する作品を選別したため，各ジャンル 390 作品を使用する．

[*1] http://taku910.github.io/mecab/
[*2] https://github.com/neologd/mecab-ipadic-neologd
[*3] https://yomou.syosetu.com/

表 1　実験結果

実験方法	正解率（%）
ベースライン	72.62
提案手法 1	79.23
提案手法 2	84.36
新聞記事	85.94

表 2　提案手法 1 による各ジャンルごとの正解率

ジャンル	正解率（%）
推理	80.55
歴史	88.82
ホラー	75.09
SF	75.64
恋愛	85.64
ファンタジー	69.73

表 3　提案手法 2 による各ジャンルごとの正解率

ジャンル	正解率（%）
推理	88.46
歴史	88.97
ホラー	76.41
SF	80.77
恋愛	94.1
ファンタジー	77.44

5.3　実験結果・考察

実験の結果，表 1 に示すように正解率は改良前と比較して 7 %向上した．正解率向上の要因として，特徴語リストを導入したことが挙げられる．表 2 より，各ジャンルの正解率を比較すると「ファンタジー」の正解率が他のジャンルより低いことがわかった．原因としては，「ファンタジー」にはファンタジー要素が強い「ハイファンタジー」と要素が低い「ローファンタジー」があり，「ローファンタジー」の作品が他のジャンルと間違えやすかった．

表 3 より，ジャンルが重複していない作品で分類をしたときの各ジャンルの正解率を比較すると，「ホラー」，「SF」，「ファンタジー」の正解率が他のジャンルより低いことがわかった．これらのジャンルの正解率が低い原因として，「ホラー」はジャンルの定義が「読者に恐怖感を与える」であり，作中の舞台や世界観などに特徴がないこと，「SF」と「ファンタジー」は両ジャンルの定義が難しく，互いの類似性が高いことが挙げられる．

また，同様の手法で行った場合，13 %の差があることから小説は新聞記事よりも分類が難しいことが確認できた．

6　おわりに

本研究では，学習データとなる小説に対する数字の置換などの前処理，特徴語の調整，学習データの選別などにより分類精度の向上を図り，実験によりその効果を確認した．しかし，新聞記事に対して改良前の分類手法を適用した結果と比較すると正解率はまだ低く，小説のジャンル推定の難しさを再確認する結果となった．今後は，深層学習など他の機械学習手法の利用，分類が難しいジャンルへの対応，マルチラベル分類問題としての分類手法の検討などを進めていきたい．

参考文献

[1] 西原弘真，白井清昭，物語テキストを対象とした登場人物の関係抽出，言語処理学会第 21 回年次大会発表論文集，pp.628-631，2015.

[2] 高田叶子，佐藤哲司，文体の類似度を考慮したオンライン小説推薦手法の提案，DEIM2017 第 9 回データ工学と情報マネジメントに関するフォーラム論文集，B5-2，2017

[3] 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，S1-5，2005.

[4] 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.