特定ジャンルの小説作成を支援するための テキスト自動分類の検討

青木 雅　　南條 浩輝　　吉見 毅彦

龍谷大学 理工学部 情報メディア学科

1　はじめに

Web の発達により，個人が音楽や絵画，文学作品といったさまざまな芸術作品を公開することができるようになった．実際に，優れた作品の製作者が世間の注目を浴びることもあり，芸術作品を創作し公開しようとする個人は増加している．音楽や絵画ではジャンルごとに基本技法があり，作者はこの技法を学ぶことによって，ジャンルからの意図しない逸脱を避けることができる．一方，文学作品においては，ジャンルごとに「こう書くべき」という技法が確立されているとはいえず，作者は製作時に自らの経験に頼ることになる．このため，著者の経験が少ない場合には，目的ジャンルから意図せずに逸脱するケースがあると考えられる．

このような背景に基づき，我々はアマチュア著者の小説作成を支援するシステムの実現に向けて研究を行う．具体的には，与えられたテキストが目的ジャンルのテキストであるかをテキスト自動分類に基づいて判定し，その判定結果を著者に提示することで小説作成支援を行うシステムについて研究を行う．本稿では，与えられたテキストが目的ジャンルのテキストかそうでないかを正しく判別するためのテキスト自動分類手法について検討を行ったので，その結果について述べる．

2　小説作成支援のためのテキスト自動分類

これまでのテキスト自動分類の研究の主な対象は論文 [1] や新聞記事など書きことばテキストであった．近年では blog[2] や小説 [3] などの話しことばに近いテキストの分類の研究も行われている．従来のテキスト自動分類の研究は主に，与えられたテキストをあらかじめ用意した多くのクラスのうち最も近いものに分類する多クラス分類の研究であり，検索サイトにおいてニュース記事や小説を適切なカテゴリに自動登録することを目的としたものである．

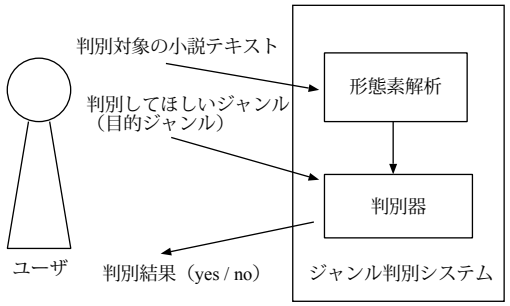

図 1: 小説作成支援システムの概要

これに対し，本研究では著者の小説作成支援を目的としているため，与えられたテキストが目的ジャンルに属するのかどうかの二値分類が重要となる．図 1 に我々が目指す小説作成支援システムの概要を示す．このシステムでは，小説テキストと判別してほしい目的ジャンルを入力として受け取り，テキスト自動分類に基づいて当該テキストが指定されたジャンルに属するかを判定する．当該ジャンルのテキストを書いたにも関わらずそうでないと判定された場合に，著者は作成した小説を本当に修正する必要があるかを判断すればよいため，この判定結果を提示することで小説作成支援が行えると考える．小説全体ではなく，より小さい単位で判別を行うことができれば，修正すべき箇所を特定して著者に伝えることができるため，よりよい小説作成支援になる．

このような背景に基づき，本稿では，当該テキストが指定されたジャンルに属するかどうかを段落単位で適切に判定する判別手法について述べる．

3　ジャンル判別手法

本研究ではサポートベクトルマシン（SVM）に基づく判別手法と言語モデルに基づく判別手法の 2 種類

の手法を検討する．

3.1 サポートベクトルマシンに基づく判別

SVM は二値分類を行うための学習機械である．特徴量が大量であっても計算時間が変わらないという特徴があるため，SVM はテキスト分類で多く用いられる [2][3][4].

本研究での SVM の特徴量は [2] に基づき，単語の出現頻度を利用する．これは各テキストから単語の出現頻度に基づく統計量を要素にもつベクトルを作成し，特徴量として用いるものである．単語の統計量には TF.IDF 値を用いる．一般的に判別対象テキスト d における単語 t の出現頻度を $tf(t,d)$，学習テキストデータ中で単語 t を含むテキストの頻度を $df(t)$，学習テキストデータの総数を N とすると，判別対象テキスト d における単語 t の TF.IDF 値 $tf.idf(t,d)$ は式（1）で与えられる．

$$tf.idf(t,d) = tf(t,d) \times \log \frac{N}{df(t)} \qquad (1)$$

ある特定のテキストに頻出し，他のテキストには出現しない単語に対してこの TF.IDF 値は大きな値となる．このように TF.IDF 値を用いることで特定のジャンルにおいて頻出する単語をより代表的な特徴として扱うことができる．

SVM に基づく判別の流れを図 2 に示す．まず，ある特定のジャンルとそれ以外に分類された小説テキストをあらかじめ用意しておき，学習データとして用いる．次に学習データの形態素解析を行って各小説テキストの単語を取得し，その出現頻度を計量する．この単語の出現頻度を基に TF.IDF 値を求めて特徴量ベクトルを生成し，学習する．判別は判別対象テキストから学習データと同様の過程で特徴量ベクトルを生成して SVM 判別器に入力することで行う．

なお，ここでの単語の単位は，Chasen-2.3.3 の形態素解析結果の「出現形」に基づいている．学習データ（後述する 4.1 節参照）における出現頻度 50 未満の単語を未知語として語彙サイズ約 36K 単語を特徴に用いた．また，SVM カーネル関数には 1 次の多項式カーネル $wx+1$ を用いた．

3.2 言語モデルに基づく判別

前述の SVM に基づく判別モデルは，単語単位での統計量のみが特徴量として用いられている [2]．このため，ジャンル特有の言い回しを特徴量に利用できていない．例えば,「ほうき＋で＋飛ぶ」といった単語列は特定ジャンル「ファンタジー」においては．頻出するが，それ以外ではあまり見られない．したがってここではジャンル特有の言い回しを特徴量として扱うために，N-gram 言語モデルを用いる．本研究では 3 単語の単語列の特徴量を利用するため，$N=3$ の単語トライグラムモデルを採用する．

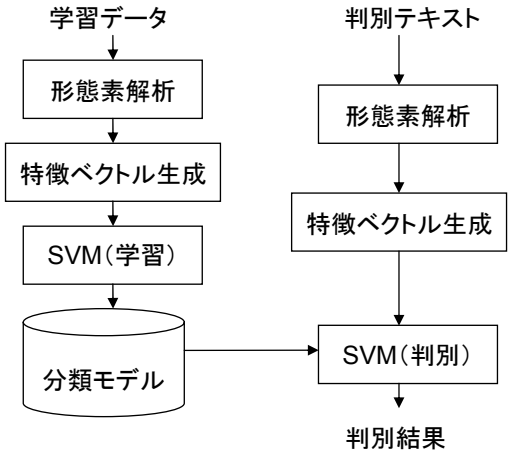

図 2: サポートベクトルマシンに基づく判別

次に，与えられたテキストとモデルの近さの尺度である単語パープレキシティを用いた判別方法について述べる．単語パープレキシティは与えられた単語列がモデルから出力される確率の 1 単語あたりの平均であり，単語列 $w_1w_2...w_n$ のパープレキシティ（PP）は式（2）で与えられる．

$$PP = (P(w_1w_2...w_n))^{-\frac{1}{n}} \qquad (2)$$

パープレキシティが低いことは，そのモデルから与えられた文字列が出現する確率が高いことを示している．すなわち，言語モデルの学習データと与えられたテキストが近いことを示している．反対にパープレキシティが高いことは学習データと与えられたテキストが離れていることを示している．複数の言語モデルが存在したとき，各モデルに対して判別対象テキストのパープレキシティを算出し，それらを比較することで与えられたテキストがどの言語モデルに近いかを判断できる．パープレキシティが低いほど言語モデルと判別対象テキストが近いため，最もパープレキシティが低い言語モデルに近いと判断できる．

言語モデルに基づく判別を図 3 に示す．学習データから目的ジャンルのテキストで学習した言語モデルと

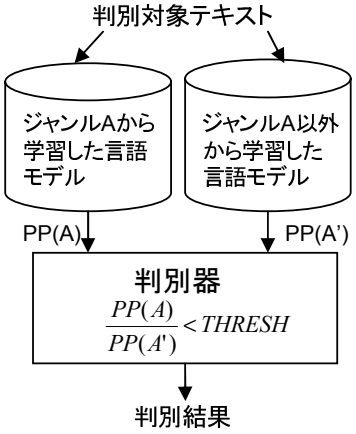

図 3: 言語モデルに基づく判別

それ以外のテキストで学習した言語モデルを作成しておき，それぞれの言語モデルで判別対象テキストのパープレキシティを算出してそれらを比較することで判別を行う（式（3））.

$$\begin{cases} \dfrac{PP(A)}{PP(A')} \leq \text{THRESH} \quad A \text{と判定} \\ \dfrac{PP(A)}{PP(A')} > \text{THRESH} \quad A \text{でないと判定} \end{cases} \tag{3}$$

ここで $PP(A)$ は目的ジャンル A で学習した言語モデルによるパープレキシティ，$PP(A')$ は目的ジャンル A 以外で学習した言語モデルによるパープレキシティである．$\dfrac{PP(A)}{PP(A')}$ が閾値以下であれば目的ジャンル A と判定し，閾値より大きければ目的ジャンル A ではないと判定する．

なお，ここでの単語の単位は SVM の実験と同一であり，Chasen-2.3.3 の形態素解析結果の「出現形」に基づいている．語彙も SVM の実験で用いるものと一致させている．

4 評価実験

4.1 評価データ

自動ジャンル分類を行うためには，学習データとしてあらかじめジャンル分類された小説テキストが必要である．本研究では，分類済みテキストとして小説検索サイト NEWVEL[5] から収集した小説テキストを用いた．NEWVEL では，ジャンルごとに小説テキストが登録されており，存在する小説のジャンルは「ファンタジー」,「SF」,「恋愛」,「ミステリー」,「現代」,「ホラー」,「文学」の 7 ジャンルと「その他」である．

表 1: 収集したデータ

ジャンル	件数
ファンタジー	2544 件
SF	423 件
恋愛	1810 件
ミステリー	164 件
現代	1288 件
ホラー	182 件
文学	145 件
その他	198 件
合計	6754 件

表 2: 分割データ件数

ファンタジー	155352 件
ファンタジー以外	186676 件
合計	342028 件

収集したデータのジャンルごとの件数を表 1 に示す．本研究では目的ジャンルかどうかの判別（二値分類）の実験を行うため，この収集データを 2 クラスにわけた．具体的には収集データにおいて最も多いジャンル「ファンタジー」とそれ以外のジャンルの 2 クラスにわけ，実験データとした．この各クラスのデータをさらに 2 分割し，学習データとテストデータとした．学習データ，テストデータの件数は共に「ファンタジー」1272 件,「ファンタジー以外」2105 件であった．

次に，段落単位での判別実験のデータについて述べる．実験データは web から収集した html 形式の小説テキストであるが，明示的な段落情報（例えば P タグ）が存在しないものも多い．そこで，本研究では原稿用紙 1 枚分に相当する 400 字で実験データを分割し，段落単位として扱った．この単位で目的ジャンルかを判断できれば小説作成支援には貢献が大きいと考えられる．分割後のテストデータの件数を表 2 に示す．データ件数は「ファンタジー」が約 15 万件,「ファンタジー以外」が約 18 万件である．

4.2 実験結果

SVM と言語モデルによる判別手法を用いて小説全体と段落単位で判別を行った．判別精度の評価には式

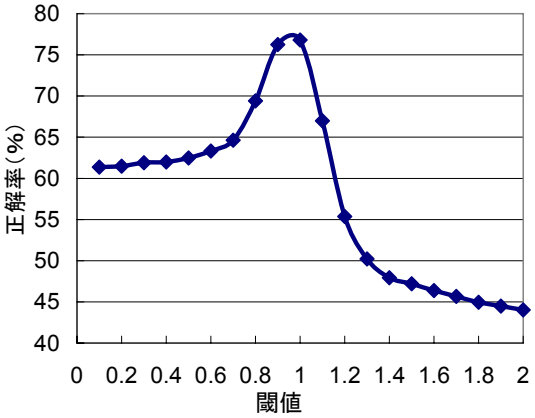

図 4: 言語モデルによる判別における閾値の影響

(4) によって求める判別正解率を用いた．

$$\text{判別正解率} = \frac{\text{システムの判別成功テキスト数}}{\text{全テキスト数}} \times 100 \quad (4)$$

はじめに言語モデルの判別における閾値パラメータ（式（3）の THRESH）の検討を行った．具体的には，閾値を 0.1 から 2 の範囲で 0.1 きざみで設定し，それぞれの場合の判別正解率を求めた．結果を図 4 に示す．なお，ここでの入力は小説全体である．閾値が 1 のときに最も判別正解率が高いことが確認できる．本稿では以降，この閾値を 1 として実験を行う．

次に，SVM と言語モデルの比較を行った．小説全体での実験結果を表 3 に，段落単位での実験結果を表 4 に示す．小説全体での判別正解率は，SVM による判別で約 72%，言語モデルによる判別で約 77%であり，言語モデルを用いた場合に高い判別正解率となった．段落単位での判別正解率は，SVM による判別では約 50%，言語モデルでの判別正解率は約 75%であった．SVM では小説全体での判別に比べて段落単位での判別では大幅に正解率が低くなっているのに対して，言語モデルでは小説全体と段落単位での正解率に大きな差がみられなかった．このことは，言語モデルで用いた単語列の特徴量は判別単位の大きさに頑健であることを示唆している．

我々が考えている小説作成支援システムは，判別された結果が著者の意図と異なる場合に，著者に該当箇所の修正についての判断を要請するものである．著者は誤りの指摘については無視すればよいため，ある程度の判別誤りは許容できると考えられる．このため，本実験で得られた 75%程度の精度でも有効な支援が行えると考える．

表 3: 小説全体での判別

判別手法	判別正解率
SVM	71.90 %
言語モデル	76.81 %

表 4: 段落単位での判別

判別手法	判別正解率
SVM	50.07 %
言語モデル	74.80 %

5 おわりに

小説作成支援を目的として，与えられた小説テキストが目的とするジャンルに属するかどうかの判別実験を行った．判別は小説全体と段落単位で行った．単語の特徴量を用いたサポートベクトルマシン（SVM）と単語列の特徴量を用いた言語モデルの 2 種類の手法に基づく判別を行ったところ，言語モデルは SVM に比べて高い判別正解率を与えることがわかった．特に段落単位での判別に言語モデルが SVM より優れていることがわかり，75% の判別精度を得ることができた．

今後は，ファンタジー以外のジャンルでの二値分類の実験を行ったうえで，作成支援の観点からの評価を行っていく予定である．

参考文献

[1] 池内淳, 安形輝, 石田栄美, 野末道子, 宮田 洋輔修一. プーリング手法を用いた学術論文の自動判別実験. 情報処理学会研究報告, 2007-DD-60.pp33-40, 2007.

[2] 池田大介, 南野朋之, 奥村学. blog の著者の性別推定. 言語処理学会第 12 回年次大会発表論文集, pp356-359, 2006.

[3] 馬場こづえ, 藤井敦, 石川哲也. 小説テキストを対象とした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, pp67-70, 2005.

[4] 高村大也, 松本裕治. 独立成分分析を用いた文書分類:svm のための素性空間再構成. 情報処理学会研究報告, 2001-NL-143-3.pp17-24, 2001.

[5] http://www.newvel.jp/.