文学小説のジャンル分類における有効な文字数の検討

情報科学科　永岡　由行　　指導教員：山村　毅

1　はじめに

図書館や書店で読みたい小説を探す場合，ジャンルを参考にすることが多い．しかし，近代文学の小説は「文学」や「著者名」などのジャンルで分類されていることが多く，「恋愛」や「歴史」などのジャンルで探したい人にとって不便である．そこで，本研究では文学小説を対象にジャンル分類を行う．

小説のジャンル分類の関連研究として，馬場ら [1] と青木ら [2] の研究がある．馬場らは，小説テキストとその小説とは異なる書籍の要旨テキストの両方を用い，分類実験を行った．その結果、小説テキストを最大で 58.2% 分類できた．青木らは，小説テキストと小説テキスト全体を 400 文字で分割したデータのそれぞれを用いて分類実験を行い，最大で 76.81% 分類できた．

2　青空文庫

青空文庫[*1]とは著作権が失われた作品のテキストを Web 上に公開している電子図書館である．無料で閲覧でき、2013 年の時点で収録作品数は 12000 件以上である [3]．本研究では，そこから収集した童話，恋愛小説，ミステリー，歴史小説，SF・ファンタジーの 5 ジャンル各 50 作品の 250 作品を実験データとする．

3　小説のジャンル分類

小説のジャンル分類では分類対象から特徴量を抽出し，それをもとに分類を行う．テキストのジャンル分類における代表的な手法に，各文章の単語の出現頻度を要素として特徴ベクトルを作成する Bag of Words(以下，BoW) と呼ばれる方法がある．BoW ではテキストの全単語を用いるため，特徴ベクトルの次元数が極めて大きくなるという問題が生じやすい．次元数が高くなると，特徴ベクトルを処理する時間が増え，次元が複雑になり，結果を得るために必要なサンプル数が多くなるため好ましくない．そこで次元数を減らすために，本研究ではジャンル分類に有効な文字数を検討する．

4　評価実験

冒頭からの文字数を 500〜20000 まで変化させながら、ジャンル分類を行い分類正解率を計算した。BoW による表現では、文章を日本語形態素解析システム MeCab で形態素解析し「単語＋品詞」の出現頻度を特徴量として用いた。また、分類には WEKA を使用し、分類器には SVM(Normalized PolyKernel、PolyKernel、RBFKernel)、Naive Bayes を使用した。評価値の算出には 10 分割交差検定を使用した。

結果を表 1 に示す．これより，SVM(PolyKernel) を使用したときに最大で 90.2% の分類正解率を得られることがわかった．しかし，冒頭からの文字数が 5000 文字以降の上昇幅は 1.2% 以下となっていることから，分類に有効な文字数は 5000 文字程度が適していると考える．また，馬場らと青木らの実験結果と比較しても高い分類結果を得られた．従って，小説のジャンル分類において文字数を制限する手法は有効だと言える．

表 1　実験結果

文字数[*2]	次元数	SVM(%)			Bayes(%)
		N-Poly[*3]	Poly	RBF	Naive
500	9232	70.4	69.2	52.8	59.6
1000	13375	72.4	75.6	52.4	69.6
2000	19004	73.6	83.2	45.2	77.6
3000	22900	72.8	86.8	40.0	85.2
4000	26020	70.0	86.4	38.0	82.4
5000	28596	75.2	89.2	36.4	82.4
6000	30797	76.0	90.4	34.8	84.0
7000	32632	76.4	91.2	34.0	85.2
8000	34337	76.8	90.8	34.4	86.4
9000	35747	75.2	92.0	33.2	83.6
10000	36911	75.2	90.8	32.8	84.8
11000	37979	76.4	89.6	31.2	85.6
12000	38960	77.2	88.8	31.6	85.2
13000	39797	78.0	88.0	32.4	85.2
14000	40574	80.0	88.0	34.0	84.0
15000	41280	81.2	86.8	31.6	84.0
16000	41937	80.4	86.4	30.8	83.6
17000	42519	80.4	85.4	32.0	85.2
18000	43116	78.4	85.2	32.8	84.4
19000	43699	76.8	84.8	33.6	84.4
20000	44197	76.4	84.8	33.6	84.4
ALL	67461	54.0	69.2	36.8	73.2

5　まとめ

本研究では，特徴ベクトルに「単語＋品詞」の出現頻度を用いた．分類実験では，4 種類の分類器を使用し，5 ジャンルに分類した．その結果，SVM(PolyKernel) を使用したときに最大で 90.2% の分類正解率を得られたが，冒頭からの文字数が 5000 文字以降の正解率の変化は小さく，文学小説のジャンル分類における有効な文字数は冒頭から 5000 文字程度であると言える．

今後の課題として，情報利得を用いて特徴ベクトルの次元数を減らすことが考えられる．これにより，更に高い分類正解率を得られる可能性がある．

参考文献

[1] 馬場こずえ, 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.

[2] 青木雅, 南條浩輝, 吉見毅彦: ”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”, 言語処理学会第 14 回年次大会発表論文集,pp627-630,2008.

[3]「「青空文庫」一転曇り空？　作品数、大幅減の懸念 :日本経済新聞」,<https://www.nikkei.com/article/DGXNASDD120G6_S3A710C1XX1000/> (参照 2018-1-18).

[*1] http://www.aozora.gr.jp/

[*2] 冒頭からの文字数

[*3] Normalized Poly