文末表現辞書を用いた文体分類とその応用

有馬　直也† 　湯本　高行†† 　礒川悌次郎†† 　上浦　尚武††

† 兵庫県立大学大学院工学研究科　〒 671–2201 兵庫県姫路市書写 2167
†† 兵庫県立大学大学院工学研究科　〒 671–2201 兵庫県姫路市書写 2167
E-mail: †ei17e001@steng.u-hyogo.ac.jp, ††{yumoto,isokawa,kamiura}@eng.u-hyogo.ac.jp

あらまし　近年，個人がインターネット上で情報を発信することが容易となっている．そのため，インターネット上には利用者の興味や関心を示す有益なデータが多くみられる．しかし，すべてのデータが有益であるとは言えず，マイニング対象を選定するには多大な労力が必要となる．そこで，トピックに対する書き手の態度に着目する．書き手の態度は文体に表れると考え，文章の文体を分類する手法を提案する．本研究では，文体の種類として敬体，常体，会話体，ネット文体を定義し，文章の文末表現に着目して分類を行う．本手法では，ニュースサイトや新聞記事などを基に文末表現辞書を構築し，この辞書を用いて文体を分類する．また，文体分類の応用として，ツイートにおける文体ごとの着目点や，文体の種類と文章の難易度との関連を調査する．

キーワード　テキストマイニング，自然言語処理，文体，文書分類

1. はじめに

近年，スマートフォンや PC などの普及により，個人がインターネット上で情報を発信することが容易となっている．そのため，インターネット上には SNS の投稿や商品レビューなど，利用者の興味や関心を示す有益なデータが多くみられる．しかし，すべてのデータが有益であるとは言えず，ジョークなど娯楽目的で投稿されたデータも多く存在する．このように，莫大な量のデータから必要となるデータのみを取り出すには多大な労力が必要となる．この問題を解決するためには，コンピュータによる自動的なデータの分別が必要となる．

そこで，本研究ではトピックに対する書き手の態度に着目する．書き手の態度は文体に表れると考え，文章の文体を分類する手法を提案する．本研究では，文体の種類として敬体，常体，会話体，ネット文体を定義し，文章の文末表現に着目して分類を行う．具体的には，各文体クラスの基としてニュースサイトや新聞記事などを用いて文末表現辞書を構築し，この辞書と同じ表現が用いられているか否かで文体を分類する．

また，文体分類の応用としてツイートにおける文体の種類ごとの意見の着目点にどのような差があるかを調査する．さらに，文体の種類と文章の難易度にどのような関連があるかを調査する．

2. 関連研究

本研究では，文章が有益か否かを文章の文体に着目して分類することを目的としている．Okamura らの研究 [1] では，ソーシャルタグを用いた機械学習により，ツイートを Positive か Negative か，さらに Positive の中でも娯楽目的の情報かそれ以外の興味や関心の高い情報かに分類を行っている．また，Nobata らの研究 [2] では，文字 N グラムやコメントの長さといった言語的特徴，さらに構文的特徴などを特徴量とした機械学習を用いてオンラインユーザーコンテンツにおける不正な言語の検出を行っている．本研究では，文末表現辞書を用いてルールベースで文章を分類するという点でこれらの研究と異なる．

また，文体に関する研究として Sato らの研究 [3] では日本語文書の難易度を推定する手法を提案している．この研究では，小学校から大学までの教科書文書をコーパスとして使用し，与えられた文書から抽出される特徴量 (ひらがなの割合，品詞など) によりその文書の難易度を数値化している．本研究では，文章の丁寧さの観点から，文体の種類として敬体，常体，会話体，ネット文体に分類する手法を提案する．

3. 文体分類器の構築

本研究での文体分類では，文章の文末表現に着目する．本手法では，ニュースサイトや新聞記事，Yahoo!知恵袋の文章を基に文末表現辞書を構築し，この辞書を用いて文体の分類を行う．

3.1 文体クラス

本研究では，文体を文章の丁寧さの観点から以下の 4 クラスを定義する．

- 敬体
- 常体
- 会話体
- ネット文体

公的な機関が情報を発信するときなど，堅い印象の文には新聞やニュースのような改まった表現が用いられることが多いと考えられる．そこで，敬体クラスは「です，ます」調となっている文章，常体クラスは「だ，である」調となっている文章と定義する．

また，ジョークなどの娯楽目的で情報を発信する場合，普段の会話で使用している話し言葉を文章で表した口語体やさらに砕けたネット特有の表現が用いられることが多いと考えられる．そこで，会話体クラスは「だよね，ですね」のように敬体クラスや常体クラスよりも柔らかい表現となっている文章と定義す

る．また，ネット文体クラスは会話体クラスよりもさらに砕けた表現となっている文章と定義する．

3.2 文末表現の定義

本研究での文体分類では，文章の文末表現に着目してクラスを決定する．本手法では，JUMAN [4] を用いて文ごとに形態素解析を行い，末尾の形態素からさかのぼり，動詞，形容詞，名詞のいずれかまでを文末表現と定義する．このとき，動詞や形容詞であれば活用形も考慮する．たとえば,「これはいい記事だ」という文では「だ」の直前が名詞「記事」であるため,「(名詞) +だ」を文末表現として抽出する．また,「これはおいしいですね」という文では「ですね」の直前が形容詞「おいしい」の基本形であるため,「(形容詞の基本形) +ですね」を文末表現として抽出する．

しかし，文章によっては上記の方法で抽出すると不必要に長い文末表現となってしまうことがある．たとえば,「食べられているらしい」や「食べられるようになった」という文から文末表現を抽出すると「(動詞の未然形) +られているらしい」,「(動詞の未然形) +られるようになった」が得られる．ここで，形態素解析器 MeCab [5] は JUMAN に比べ，文章をさらに細かく形態素に分割するという特徴がある．たとえば,「している」という表現は JUMAN では「して (動詞)/いる (接尾辞)」となるが MeCab では「し (動詞)/て (助詞)/いる (動詞)」と解析される．そこで，MeCab で同文を解析した際，以下の条件を満たす場合にその部分を区切りと設定し，文末表現を抽出する．

- JUMAN で接尾辞，MeCab で接続助詞となった場合
- JUMAN で助動詞，MeCab で名詞となった場合

このルールに従って前述の「食べられているらしい」という文から文末表現を抽出すると，JUMAN では「られて」が接尾辞，MeCab では「て」が接続助詞となるため「(動詞の連用形) +いるらしい」を文末表現として抽出する．同様に,「食べられるようになった」という文から文末表現を抽出すると，JUMAN では「ように」が助動詞，MeCab では「よう」が名詞となるため「(名詞) +になった」を文末表現として抽出する．

3.3 文末表現辞書の構築

本研究では各クラスの基となるデータとして，敬体クラスはニュースサイトの記事を，常体クラスは新聞記事を，会話体クラスは Yahoo!知恵袋を，ネット文体クラスは，はてなブックマークを使用する．

また，本手法では文章の文体を分類するため，ネット文体クラスを除く 3 つのクラスで文末表現辞書を構築する．このとき辞書には，クラスの基となる文章に対して形態素解析を行って得られた文末表現を使用する．しかし，新聞記事には会話文やコラムが存在するため,「〜です」や「〜だね」のように常体クラスに属さない表現が抽出されることがある．そこで，敬体辞書および会話体辞書を用いて常体辞書内の常体表現以外を除去する．具体的には，常体辞書内の表現のうち，敬体辞書や会話体辞書にも共通してみられる表現を取り除く．ただし，常体辞書内でみられる会話体表現は頻度が高くないため，会話体表現を除去する場合は常体辞書で頻度の低い表現のみを対象とする．

また，文末表現辞書を用いた分類では，辞書内の表現と完全一致する候補が見つからない場合は末尾からの部分一致を考慮するため，辞書のデータ構造にはトライ木を用いる．この詳細は 3.4 節で述べる．

3.4 文末表現辞書を用いた文体分類

文体分類の概要を図 1 に示す．文体分類では 3.3 節で述べた辞書を用いる．まず，各文から文末表現を抽出し，それらを各辞書と照らし合わせることで文体クラスを決定する．この詳細を以下に示す．

まず，各文から文末表現を抽出する (図 1 の 1)．ただし，文末が “」” で終わる文は対象外とする．このとき，どの文からも文末表現が抽出できない場合はネット文体クラスと分類する．

次に，抽出した文末表現を各辞書と照らし合わせる (図 1 の 2)．このとき，動詞や形容詞の原型あるいは名詞であれば常体辞書に含まれる表現とみなす．また，抽出した文末表現がどの辞書の表現とも完全一致しないときは，各辞書の表現と文末から部分一致するかを比較し，部分一致すればその辞書に含まれる表現とみなす．ただし，部分一致を考慮する文字列の長さは 2 文字以上とする．

最後に，各辞書と照らし合わせた結果より，次のルールで文体クラスを決定する (図 1 の 3)．

(i). 全ての文末表現が敬体辞書に存在するとき敬体クラスに分類

(ii). 全ての文末表現が常体辞書に存在するとき常体クラスに分類

(iii). 敬体辞書だけに含まれる文末表現と常体辞書だけに含まれる文末表現が混合して存在するとき敬体クラスと常体クラスの両方に分類

(iv). 会話体辞書にしか存在しない文末表現が 1 つでも存在するとき会話体クラスに分類

(v). 上記以外のときネット文体クラスに分類

文末表現を抽出する際に砕けた文章であると JUMAN で正しく品詞を解析できず，誤って名詞や動詞となることがある．たとえば,「おもしれー」を JUMAN で解析すると「れー」が名詞「礼」と解析される．そこで，文末が動詞や形容詞の原型あるいは名詞であるとき MeCab 辞書を併用し，その解析結果が同じである場合に正しい品詞であると判断する．ただし，MeCab では JUMAN に比べ，細かく形態素ごとに分割されてしまうので，MeCab での解析結果で名詞あるいは動詞の後に助動詞が続く場合も，JUMAN で得られた品詞が正しいと判断する．たとえば,「最高だった」を解析すると JUMAN では「最高だった (形容詞)」となるが，MeCab では「最高 (名詞)/だっ(助動詞)/た (助動詞)」となる．このような場合も，JUMAN と MeCab の解析結果が一致したとみなす．

4. 評価実験

ネット文体を除く 3 つのクラスの文末表現辞書を用いて文体を分類する．分類精度の評価は適合率，再現率の調和平均であ

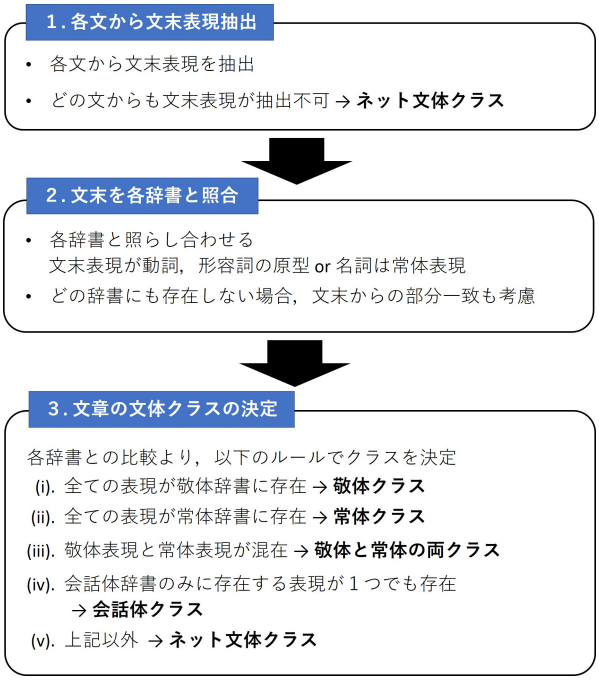

図 1　文体分類の流れ

表 1　データセット

	辞書構築用データ	テストデータ
敬体クラス	826	100
常体クラス	4445	120
会話体クラス	10000	103
ネット文体クラス		100

表 2　文体分類の結果

	適合率	再現率	F 値
敬体クラス	0.990	0.990	0.990
常体クラス	0.878	0.658	0.752
会話体クラス	0.662	0.932	0.774
ネット文体クラス	0.592	0.520	0.553

表 3　結果の詳細

		分類結果			
		敬体	常体	会話体	ネット文体
正解	敬体	99	0	1	0
	常体	0	79	8	33
	会話体	1	3	96	3
	ネット文体体	0	8	40	52

る F 値によって行う．この F 値が大きいほど，分類精度が良いといえる．これらの評価指標の定義を以下に示す．

適合率（Precision）：分類器の出力結果のうち，実際に正解のデータである割合

$$\text{適合率} = \frac{\text{正しく正例と分類できた数}}{\text{分類器が正例と分類した数}} \quad (1)$$

再現率（Recall）：全ての正解データのうち，分類器が正解として出力した割合

$$\text{再現率} = \frac{\text{正しく正例と分類できた数}}{\text{正解データの正例の数}} \quad (2)$$

F 値：適合率と再現率の調和平均

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \quad (3)$$

4.1　データセット

実験で使用したデータを表 1 に示す．本実験では辞書を構築するための基となる文章として，敬体クラスは NHK NEWS WEB (注1) で 2017 年 10 月 18 日～24 日に記載されたニュース記事を，常体クラスは 2014 年 1 月 1 日～20 日の毎日新聞の記事を，会話体クラスは 2004 年 4 月 2 日～5 月 29 日に投稿された Yahoo!知恵袋の質問文および回答文を使用した．

また，テストデータとして敬体クラスは NHK NEWS WEB で 2017 年 10 月 24 日～25 日に記載されたニュース記事を，常体クラスは 2014 年 4 月 17 日～18 日の毎日新聞の記事を，会話体クラスは 2004 年 4 月 5 日～5 月 4 日に投稿された Yahoo!知恵袋の質問文および回答文を，ネット文体クラスは 2017 年 8 月 31 日に投稿されたはてなブックマークのコメントを使用した．

(注1)：http://www3.nhk.or.jp/news/

上記のデータは著者のうち 1 名が収集し，さらにテストデータには，収集したデータからそのクラスの文体に適した文章となっているデータのみを用いた．

4.2　文体分類の評価

表 1 に示したデータを用いて，3.3 節で述べた手法により辞書を構築した．このとき，常体辞書に含まれる敬体表現を除去するため，敬体辞書に含まれるすべての表現を常体辞書から取り除いた．さらに，会話体表現を除去するため，会話体辞書に含まれ，かつ常体辞書で頻度が 20 以下の表現を常体辞書から取り除いた．これにより，敬体辞書には 43 表現が，常体辞書には 589 表現が，会話体辞書には 2091 表現が得られた．

この辞書を用いてテストデータを分類したときの分類精度を表 2 および表 3 に示す．表 2 より，敬体，常体，会話体クラスでは F 値が 0.7 以上と高い精度であることがわかる．また，表 3 より常体クラスではネット文体クラスへの誤分類が多く，ネット文体クラスでは会話体クラスへの誤分類が多くみられる．前者は，新聞記事で文末が「～という」といった終わり方をしていることがあり，これは JUMAN で「いう」が動詞の基本形，MeCab で助詞と解析結果が異なっていることが原因であった．また，後者は「～だわ」や「～なぁ」といった表現が会話体辞書にも存在していることが原因であった．

5.　文体分類の応用

文体分類の応用として，まず，トピックに対する書き手の態度が文体に表れるとすると，文体ごとの意見の着目点に違いがみられると考えられる．そこで，本手法の文体分類を用いてツイートにおける文体ごとの意見の着目点を調査する．また，難しい内容は硬い文体で書かれると考えると，本手法の文体分類により簡易的な文章難易度の推定が可能であると考えられる．そこで，文体の種類と文章の難易度との関連を調査する．

5.1 ツイートにおける文体ごとの意見の着目点の調査

本実験では，ツイートにおける文体の種類ごとの意見の着目点にどのような差があるかを調査する．具体的には，指定したキーワードを含むツイートを収集し，ツイートの文体を文末表現辞書を用いて自動的に分類する．文体クラスの決定後，各文体ごとにツイートから名詞を抽出する．本実験では，トピックによる違いを調査するため，キーワードとして「ダイエット」,「EU 離脱」,「サッカー」,「iPhone」,「大統領」のいずれかを含むツイートを用いる．

本実験には，2014 年 11 月〜2016 年 9 月の間に投稿されたツイートをダイエットに関しては 1070 件，EU 離脱に関しては 1757 件，サッカーに関しては 2580 件，iPhone に関しては 2310 件，大統領に関しては 2230 件使用した．このツイートを用いて，文体ごとに名詞を抽出したときの結果上位 10 件を表 4〜表 8 に示す．ただし，ツイートの多くは砕けた表現が多く，文体によってデータ数が少なくなる場合がある．そのため，使用したデータ数は文体によって異なっている．また表 4〜表 8 において，トピック語は除外し，さらにトピックと関係のない語をストップワードとするため，実験に用いるトピックと関連の薄い「君の名は」,「野球」,「パナマ文書」,「イチロー」,「オリンピック エンブレム」をキーワードとして含む各 500〜1000 件のツイートから名詞を抽出し，さらに全トピックに共通して出現し，かつ頻度が 3 以上の名詞をストップワードとして使用した．

表 4 より，ダイエットに関する文体ごとの名詞を比較すると，どの文体も上位の名詞には「脂肪」や「運動」といった共通する語がみられる．また，表 5〜表 8 より他のトピックでも同様に，上位 5 件程度ではすべての文体で共通する語が使用されていることがわかる．しかし，表 5 や表 6 に示した語より頻度が下位の語を比較すると，文体ごとに使用されている語に違いがみられる．表 5，表 6 における文体ごとの語の順位を表 9，表 10 に示す．

表 9 より,「後悔」という語は常体で高い順位であるが，会話体とネット文体では順位が低く，さらに敬体ではその後が存在していないことがわかる．表 10 においても同様に,「シュート」という語は敬体では順位が高いが，他の文体では順位が低いことがわかる．これらの結果より，頻度が上位の語については文体ごとに使用されている語に差はみられないが，それより下位の語では文体によって使われない語があるなど，トピックによっては差があると考えられる．

また，極性の観点から文体ごとの意見極性にどのような違いがあるかを調査するため，文体ごとのツイートをさらに Positive か Negative か，あるいは Neutral かに分類し，文体ごとの意見極性の偏りを比較した．このときの結果を表 11〜表 15 に示す．このとき，極性分類には Google Natural Language API [6] を用いた．

表 11〜表 15 より，表 14 の iPhone に関するツイート以外はどの文体でも Positive のツイートが最も多くみられる．しかし，表 14 では常体，会話体，ネット文体に Negative の文が最も多くみられる．この結果より，トピックによっては文体ごと

表 4　ダイエットに関する結果

敬体 (300 件)		常体 (250 件)	
名詞	頻度	名詞	頻度
脂肪	85	自分	28
効果	71	あなた	19
カロリー	60	もの	19
体	57	脂肪	19
代謝	51	運動	16
運動	41	成功	16
筋肉	34	体	15
消費	33	体重	15
ため	32	カロリー	15
食事	29	グルメ	15

会話体 (350 件)		ネット文体 (170 件)	
名詞	頻度	名詞	頻度
脂肪	71	運動	19
効果	47	脂肪	18
ため	43	体重	15
食事	41	効果	14
運動	39	カロリー	13
カロリー	39	自分	13
もの	39	体	13
代謝	36	もの	11
お腹	32	消費	11
体重	32	糖質	10

表 5　EU 離脱に関する結果

敬体 (57 件)		常体 (600 件)	
名詞	頻度	名詞	頻度
イギリス	35	イギリス	293
国民	9	英国	79
投票	7	投票	77
英国	7	日本	59
今日	6	国民	57
結果	6	世界	53
アメリカ	5	影響	42
日本	5	英	42
今後	4	経済	39
気	4	問題	34

会話体 (600 件)		ネット文体 (500 件)	
名詞	頻度	名詞	頻度
イギリス	312	イギリス	215
英国	75	英国	72
投票	75	投票	45
日本	55	日本	41
影響	48	国民	37
経済	47	経済	35
国民	47	英	29
問題	43	派	29
国	41	国	25
派	35	影響	22

表 6　サッカーに関する結果

敬体 (180 件)		常体 (800 件)	
名詞	頻度	名詞	頻度
試合	25	部	151
野球	20	選手	87
選手	19	ボール	44
部	14	日本	40
日本	12	俺	38
もの	11	時	37
スポーツ	11	代表	37
相手	11	野球	35
チーム	10	チーム	33
高校	10	リーグ	30

会話体 (800 件)		ネット文体 (800 件)	
名詞	頻度	名詞	頻度
選手	76	部	129
部	73	俺	55
試合	65	野球	52
野球	52	日本	51
ボール	49	試合	49
俺	40	ボール	47
僕	35	選手	43
日本	31	スポーツ	42
チーム	29	代表	37
今日	29	湘南	37

表 7　iPhone に関する結果

敬体 (110 件)		常体 (800 件)	
名詞	頻度	名詞	頻度
修理	14	ケース	67
機種	7	アップデート	56
画面	7	画面	44
返信	5	音	27
今日	5	ボタン	25
店	5	充電	25
垢	5	予約	24
シャッター	5	ホーム	23
音	5	アプリ	20
在庫	4	イヤホン	20

会話体 (600 件)		ネット文体 (800 件)	
名詞	頻度	名詞	頻度
ケース	45	ケース	76
アプリ	38	アップデート	42
画面	36	画面	39
音	29	音	34
アップデート	25	時	31
イヤホン	22	充電	29
修理	20	電話	24
充電	19	俺	23
気	19	スマホ	21
時	18	アプリ	20

表 8　大統領に関する結果

敬体 (130 件)		常体 (700 件)	
名詞	頻度	名詞	頻度
アメリカ	38	アメリカ	129
オバマ	22	日本	72
日本	18	米	68
米	14	選	66
米国	13	オバマ	53
候補	13	トランプ	48
世界	13	国	43
選挙	12	米国	36
政府	12	リンカーン	36
トランプ	11	候補	34

会話体 (700 件)		ネット文体 (700 件)	
名詞	頻度	名詞	頻度
アメリカ	144	アメリカ	110
トランプ	104	日本	80
日本	77	トランプ	68
選	67	選	55
オバマ	65	米	55
国	57	国	54
選挙	54	世界	52
候補	48	オバマ	36
クリントン	44	首相	34
世界	41	米国	33

表 9　EU 離脱に関する頻出語の文体ごとの順位

	敬体	常体	会話体	ネット文体
イギリス	1 位	1 位	1 位	1 位
投票	3 位	3 位	2 位	3 位
影響	9 位	7 位	5 位	10 位
文化	12 位	302 位	-	-
後悔	-	21 位	112 位	68 位
危機	69 位	25 位	112 位	68 位

表 10　サッカーに関する頻出語の文体ごとの順位

	敬体	常体	会話体	ネット文体
試合	1 位	10 位	3 位	5 位
選手	3 位	2 位	3 位	5 位
チーム	9 位	9 位	9 位	12 位
相手	6 位	69 位	50 位	113 位
シュート	17 位	88 位	119 位	481 位
部員	-	142 位	29 位	94 位

表 11　意見極性の偏りの結果 (ダイエット)

	Positive	Negative	Neutral
敬体	80.0%	11.1%	8.9%
常体	64.8%	22.0%	13.2%
会話体	75.7%	12.0%	12.3%
ネット文体	68.8%	13.0%	18.2%

表 12　意見極性の偏りの結果 (EU 離脱)

	Positive	Negative	Neutral
敬体	52.6%	21.1%	26.3%
常体	44.0%	32.8%	23.2%
会話体	42.5%	35.7%	21.8%
ネット文体	38.2%	37.2%	24.6%

表 13　意見極性の偏りの結果 (サッカー)

	Positive	Negative	Neutral
敬体	81.7%	11.1%	7.2%
常体	67.6%	20.8%	11.6%
会話体	64.5%	20.8%	14.7%
ネット文体	66.8%	19.7%	13.5%

表 14　意見極性の偏りの結果 (iPhone)

	Positive	Negative	Neutral
敬体	50.0%	34.5%	15.5%
常体	39.4%	47.6%	13.0%
会話体	34.5%	48.5%	17.0%
ネット文体	37.6%	44.6%	17.8%

表 15　意見極性の偏りの結果 (大統領)

	Positive	Negative	Neutral
敬体	66.9%	17.7%	15.4%
常体	51.6%	29.1%	19.3%
会話体	44.9%	36.7%	18.4%
ネット文体	50.9%	29.3%	19.8%

に意見極性の傾向が異なるということがわかる．また，敬体クラスの意見極性はどのトピックにおいても Positive の文が最も多くみられる．この原因として,「サッカーしてきます」のように書き手の感情が述べられていない文においても Positive と分類されていることが考えられる．さらに，敬体には「〜ができます」のように読み手へのアドバイスや宣伝目的で書かれたツイートも多くみられ，このようなツイートは読み手に良い印象を与えるために Positive な内容で書かれることが多い．このように，アドバイスや宣伝目的のツイートが敬体クラスに多いということも，敬体クラスで Positive のツイートが多くなる原因の一つであると考えられる．

5.2　文体の種類と文章の難易度の関連の調査

本実験では，文体の種類と文章の難易度にどのような関連があるかを調査する．具体的には,「EU 離脱」や「オリンピック」など複数のトピックに関するブログを収集し，その文体を文末表現辞書を用いて自動的に分類する．このとき，収集するブログには Ameba ブログ (注2) を用いる．また，さらにその文章の難易度を日本語文章難易度判別システム (注3) [7] [8] を用いて判別することで，文体の種類と文章の難易度との関連を比較する．

本実験に使用するデータとして 2007 年 4 月〜2018 年 1 月に投稿され，さらに複数のトピックに関するブログを収集した．この結果，敬体クラスは 45 件，常体クラスは 46 件，会話体クラスは 47 件，ネット文体クラスは 42 件集まった．このときの各トピックごとのデータを表 16 に示す．また，この文の難易度を日本語文章難易度判別システムを用いて判別したときの結果を図 2，表 17 に示す．

図 2，表 17 より，文体ごとに文章の難易度を比較すると敬体と常体では「ややむずかしい」,「むずかしい」の難易度の文章

(注2) : https://official.ameba.jp
(注3) : http://jreadability.net

表 16　トピックごとのデータ

	敬体	常体	会話体	ネット文体
EU 離脱	11	11	20	13
自動運転	4	2	2	3
藤井四段	4	2	7	5
ips 細胞	1	4	1	1
弾道ミサイル	8	9	4	6
オリンピック	2	1	2	3
オスプレイ	2	4	1	4
パナマ文書	6	6	6	4
トランプ大統領	7	7	4	3

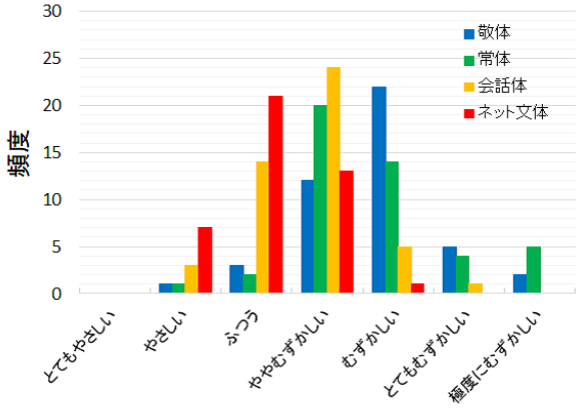

図 2　文体の種類と難易度の比較結果

表 17　文体の種類と難易度の比較結果の詳細

	敬体	常体	会話体	ネット文体
とてもやさしい	0	0	0	0
やさしい	1	1	3	7
ふつう	3	2	14	21
ややむずかしい	12	20	24	13
むずかしい	22	14	5	1
とてもむずかしい	5	4	1	0
極度にむずかしい	2	5	0	0

が多く，会話体やネット文体では「ふつう」,「ややむずかしい」の難易度の文章が多くみられる．また，日本語文章難易度判別システムでは文の難易度を数値化したスコアが得られる．このスコアは，小さい値ほど難易度が高いことを表す．このスコアを各トピックごとに求めたときの結果を図 3 に示す．

図 3 より，ほとんどのトピックにおいて敬体と常体は，会話体とネット文体よりもスコアが小さくなっていることがわかる．また，このスコアをトピックに関係なく文体ごとに合計し，平均を求めると敬体クラスは 2.25，常体クラスは 2.18，会話体クラスは 3.27，ネット文体クラスは 3.78 となった．この結果より，敬体と常体は文章の難易度が高い文が多く，会話体とネット文体では易しい文が多いと考えられる．したがって，本手法の文末表現辞書を用いて文体を分類することにより，簡易的な文章難易度の判別が可能であると考えられる．

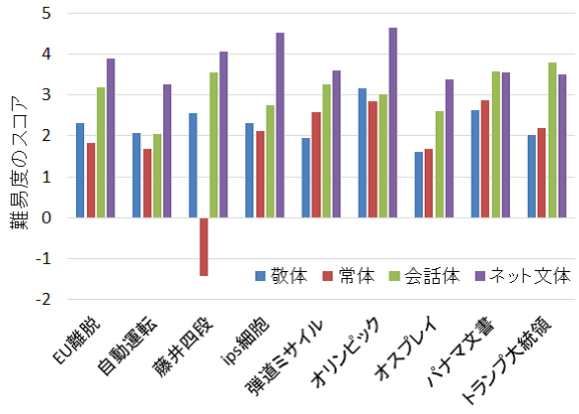

図 3　トピックごとの難易度の比較結果

6. おわりに

本研究では，文章の文末表現に着目し，文体を文章の丁寧さの観点から敬体，常体，会話体，ネット文体に分類する手法を提案した．文体分類の方法は，各文体クラスの基となる文章から文末表現辞書を構築し，文書の各文から得られた文末表現をこの辞書と照らし合わせ，ルールベースで文体を分類した．この結果，文体分類の実験では敬体，常体，会話体の 3 クラスで F 値が 0.7 以上と高い精度が得られた．

また，文体分類の応用として，文体ごとの意見の着目点について調査した．実験では，指定したトピックに関するツイートから名詞を抽出し，文体ごとに使用されている語を比較すると，上位の語は文体に関係なく同じ語がみられた．しかし，下位の語を比較すると他の文体ではほとんど使用されていない語がみられることもあった．この結果から，頻度が下位の語は，トピックによっては文体ごとに違いがあることがわかった．また，意見極性の観点から文体ごとの意見極性の違いを調査したところ，トピックによっては文体ごとに意見極性が異なることがあり，さらに敬体クラスはトピックによらず Positive のツイートが多くなる傾向があるとわかった．

さらに，文体分類の応用として，文体の種類と文の難易度との関連を調査した．実験では，文体ごとに文の難易度を判定すると，ほとんどのトピックで敬体と常体の文の難易度が会話体とネット文体の文の難易度よりも高いという結果が得られた．この結果から，本手法で提案した文体分類により簡易的な文章難易度判別が可能であるとわかった．

今後の課題として，文末表現辞書には抽出したすべての表現を使用しているため，その文体に不適切な表現もわずかに含まれていることがある．これを防ぐため，頻度を考慮し，辞書に用いる表現に閾値を設定する必要があると考えられる．また，文体を決定する際に敬体辞書や常体辞書に含まれない表現が 1 つでもあると，会話体やネット文体に分類している．そのため，このルールでは形態素解析でミスがあると誤分類を起こしてしまうことがある．これを防ぐため，敬体辞書や常体辞書に含まれない表現の出現した割合から，その文体を決定する必要があると考えられる．

謝　辞

本研究の一部は，平成 29 年度科研費基盤研究 (C)(17K00429) によるものである．また，本研究では科研費（課題番号 25370573）の成果物である「日本語文章難易度判別システム」を利用した．

文　献

[1] Yasuyuki Okamura, Takayuki Yumoto, Manabu Nii, Naotake Kamiura. “Sentiment Estimation of Tweets by Learning Social Bookmark Data”, International Journal on WWW/Internet, Vol.14, No.1, pp.15-27, 2016.

[2] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. “Abusive language detection in online user content”, In Proceedings of the 25th International Conference on World Wide Web, pp.145-153, 2016.

[3] Satoshi Sato, Suguru Matsuyoshi, Yohsuke Kondoh. “Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus”, In Proceedings of the 6th International Conference on Language Resources and Evaluation (LRCE), pp.654-660, 2008.

[4] JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[5] MeCab, http://taku910.github.io/mecab/

[6] Google Natural Language API, https://cloud.google.com/natural-language/

[7] 李 在鎬. “日本語教育のための文章難易度研究”, 早稲田日本語教育学, Vol.21, pp.1-16, 2016.

[8] Yoichiro Hasebe, Jae-Ho Lee. “Introducing a Readability Evaluation System for Japanese Language Education”, Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J), pp.19-22, 2015.