ランダムフォレストを用いた 政治テキストの社会言語学的分析のための著者固有表現抽出

鈴木崇史
東京大学大学院学際情報学府

影浦峡
東京大学大学院教育学研究科

本研究では，総理大臣の政治スタイルや個人的特徴を文体的特徴から分析するために，ランダムフォレストを用いて，テキスト分類に有効な変数の抽出を行なった．文体分析のための基本的特徴量であり，また，文章のモダリティを構成する助詞・助動詞分布を用いてテキストを分類したところ，比較的高い分類精度・再現率を得，また，著者固有の助詞・助動詞を抽出することができた．既存の政治テキストの社会言語学的分析では，分析対象とする表現を分析者が主観的に選択していたが，これに対し本研究は，テキスト分類を応用することで，その著者に固有の文体的特徴を抽出し，まさにその著者固有の政治スタイルや個人的特徴の分析が可能となること示すものである．

1　はじめに

本研究では，政治テキストの社会言語学的分析のために，ランダムフォレストを用いた著者固有表現抽出を行なう．政治リーダーの文体的特徴を分析する研究類型 [1,9] の，大きな目的の一つは，政治リーダーの文体的特徴から彼ら／彼女らの政治スタイルや個人的特徴を明らかにすることにある．しかし，この分野の既存研究では，分析者が独自の関心に基づいて分析表現を選択しており，その点で時に，解釈の恣意性や矛盾した結論を導く可能性があった．すなわち，既存研究は，政治リーダーの文体的特徴から政治的スタイルや個人的特徴を論じているものの，その文体的特徴が，まさに，その政治リーダーに特有のものであり，分析結果が，まさにその政治リーダー独自の政治的スタイルや個人的特徴を示しているのかは，明確ではなかった．

この点を解決するために，本研究では，ランダムフォレストを用いて，テキスト分類に有効な変数の抽出を行なう．政治リーダーごとに演説を分類し，分類に寄与の大きい変数を抽出することで，その政治リーダー固有の表現を抽出する．これを分析することで，まさに，その政治リーダーがもつ文体的特徴から，その政治リーダーの政治スタイルを分析することが可能となる．また，これにより，ある政治リーダーの文体が，特定表現の出現が少ないことによって特徴づけられる場合でも，その表現を分析対象とすることが可能になる．本研究で適用する手法は，政治テキスト以外の様々なテキストの社会言語学的分析に有効であり，潜在的に重要な文体論の応用分野であるとされながら [4]，まだ実際の展開事例は少ない，コンピュータ社会言語学の領域を切り開くものである．事例として，中曽根と小泉の国会演説をとりあげる．両者は，戦後日本政治の中で，もっとも強いリーダーシップをもっていた二人の政治リーダーであり，また演説，政治言語への関心も高く [9,16]，文体的特徴を分析する価値がもっとも大きいと考えられるためである．

2　データと手法

データベース「世界と日本」[*1] より中曽根，小泉，その他の総理大臣 (1980 年から 1989 年および，1989 年から 2006 年) の国会演説をダウンロードし，ChaSen [13] により形態素解析を適用する．表 1 は，それぞれの演説本数，延べ語数，異なり語数，分析対象とする助詞／助動詞の延べ語数，異なり語数である．

中曽根の演説と 1980 年から 1989 年に行なわれたその他の総理大臣の演説，小泉の演説と 1989 年から 2006 年に行なわれたその他の総理大臣の演説

[*1] www.ioc.u-tokyo.ac.jp/~worldjpn

表 1　コーパスの基礎データ

		全語彙		助詞		助動詞*	
	演説	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	異なり語数
中曽根	10	47419	3390	13851	70	4241	33
1980-1989	9	35979	2862	10656	64	3434	29
小泉	11	46979	3908	13677	62	3680	35
1989-2006	31	149234	5774	44712	80	13292	43

* ステミング前の値.

を，ランダムフォレスト [5] を用いて分類する．ランダムフォレストは決定木ベースであり，分類に有効な変数の重要度を推定することに，特徴がある [6]．分類結果の妥当性を確認したのち，分類に有効な変数 (ジニ係数上位) を抽出する．抽出された表現を分析することで，その政治リーダー固有の文体的特徴を分析し，まさにその政治リーダーがもつ政治スタイルや個人的特徴を分析することが可能となる．関連研究として，Argamon ら [2] は，性差によるテキスト分類と固有表現の抽出を SVM と Information Gain を用いて行なっているが，ランダムフォレストは直接的に分類に用いる変数を返すため，より適切である (c.f., [6])．ランダムサンプリングする変数は，Breiman に従い，変数の正の平方根，ランダムサンプリングの回数は 1000 回とし，ランダムサンプリングの 2/3 を学習に用い，残りの 1/3 を評価に用いる．鈴木・影浦 [15] により，時代による，総理演説の文体的特徴量の変化が指摘されているため，同時代の総理大臣を比較対象とし，同時代の範囲は，政治学の標準的な時代区分 [12, 17] を参考にして，これを選択する．ランダムフォレストは，多くの大規模データ分類への有効性が指摘されているものの，テキスト分類への応用事例はいまだわずかしかないため [11]，その有効性を確認することは，広く文体論に意義をもつ．

分類に使用する特徴量として，助詞および助動詞の相対頻度分布を用いる．特徴量としては，まず，一定以上の分類性能が期待されるものを選択する必要がある．一般に著者推定では，内容語に依存しない文章の機能的特徴量が用いられ，助詞・助動詞分布はその中でも基本的なものである [7]．日本語の著者推定で，助詞分布 (品詞タグ 1 階層) の有効性は，既に指摘されており [10]，また，総理演説の助詞の使用と助動詞の使用には，相関が指摘されているため [15]，助動詞分布も分類に有効であると推察される．

同時に，特徴量として，政治テキストの社会言語学的分析に有効なものを選択する必要がある．助詞・助動詞は文章のモダリティをあらわし，話し手や書き手の意図を知る手がかりとなる [14]，また，いくつかの助詞，助動詞は，実際に政治テキストの社会言語学的分析に用いられている [9] ため，これらの観点からも適切な特徴量である．

ステミング，品詞タグの階層，および複数の特徴量の組み合わせは，分類結果へ影響を与えると予想されるため，助詞 (1 階層／2 階層)，助動詞 (ステミングあり／なし)，およびそれぞれの組み合わせの計 8 通りの特徴量を比較する．ChaSen の品詞タグにより助詞，助動詞を抽出し，漢字とかなという点でのみ区別される 8 組 (16 語) の表記ゆれを統一する．助詞の異なり語数は 105, 助動詞 (ステミング前) の異なり語数は 58 である．

3　結果と考察

表 2 は，分類の精度，再現率，F_1 値を示している．中曽根よりも小泉の方が高い分類結果を示している．これは，学習する演説の本数が多いこと，また，図 1，図 2 に示されている様に，小泉の方が中曽根よりも個性的な助詞／助動詞の使用傾向をもっていたことによる．両者の分類結果は，文体的特徴の男女差を検討した関連研究の基準 (70% 以上) [3] と比べても，比較的高い分類性能と判断でき，これは，ランダムフォレストの文体論への有効性を示すものである.*2 著者推定分野での先行研究で用いら

*2 実際には，本研究で用いたテキストと特徴量に関しては，SVM の方が分類性能がよかったが，著者推定においては，分類機の性能は，特徴量に依存することが知られているため [8]，代替的手法の検討はそれ自体として意味をもつ．

表 2　精度／再現率／ F_1 値

		中曽根			小泉		
		精度	再現率	F_1 値*	精度	再現率	F_1 値*
助詞	(1 階層)	70.0	70.0	70.0	50.0	18.2	26.7
助詞	(2 階層)	70.0	70.0	70.0	83.3	45.5	58.8
助動詞	(ステミングなし)	66.7	60.0	63.2	100.0	81.8	90.0
助動詞	(ステミングあり)	77.8	70.0	73.7	100.0	90.9	95.2
組み合わせ	(1 階層, ステミングなし)	54.5	60.0	57.1	100.0	81.8	90.0
組み合わせ	(2 階層, ステミングなし)	61.5	80.0	69.6	100.0	81.8	90.0
組み合わせ	(1 階層, ステミングあり)	70.0	70.0	70.0	100.0	90.9	95.2
組み合わせ	(2 階層, ステミングあり)	77.8	70.0	73.7	100.0	90.9	95.2

* F_β 値 $= \frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}$ (P: 精度, R: 再現率)

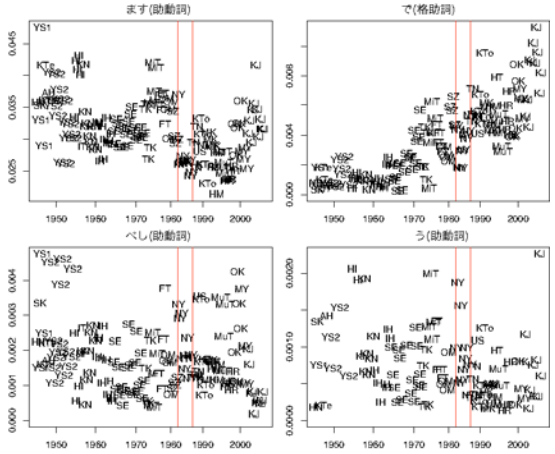

図 1　中曽根の演説に特徴的な 4 語の推移 (中曽根:NY)

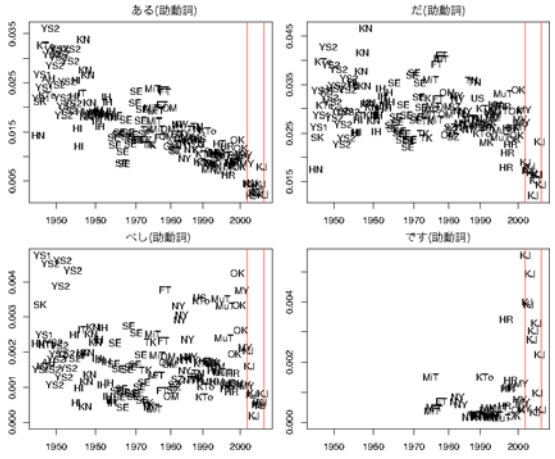

図 2　小泉の演説に特徴的な 4 語の推移 (小泉:KJ)

れていた助詞分布 [10] よりも，助動詞分布，組み合わせの方が分類性能がよく，また，ステミングとより深い階層を用いることで，分類性能を向上させ得ることを示している．

図 3 は，助詞 (2 階層) と助動詞 (ステミング後) の組み合わせを用いた，それぞれの分類におけるジニ係数上位 10 語である．ウィルコクソン順位和検定を適用し，中曽根／小泉のこれらの助詞・助動詞使用が，有意に同時代の総理大臣と異なるかを検定した. 多くの助詞・助動詞が中曽根／小泉において，有意に多く，もしくは，少なく使用されていることが示されている．

個々の助詞・助動詞の使用についてより詳しく検討するために，図 1，図 2 に，中曽根・小泉に特徴的な変数上位 4 語ずつの相対頻度を，1945 年から 2006 年までの全総理演説に対してプロットした．図からは，中曽根の「ます」「で」や小泉の「ある」「だ」「べし」のように，低頻度によって特徴づけられる表現も抽出されていることが確認される．これらの語彙は，丁寧表現 (ます，ある，です)，断定 (だ，です)，規範 (べし)，意図 (う) などを含み，それぞれの総理大臣の政治スタイルの説明に有用である．例えば，「ある」はフォーマルな演説，国会演説に特有の表現であり [9]，小泉において「ある」が少ないことは，小泉が親しみやすい政治スタイルを意図したことを反映していると解釈できる．以上から，政治テキストの社会言語学的分析のために，ランダムフォレストを用いた著者固有表現の抽出が有効であると結論づけられる．

4　おわりに

本研究では，政治テキストの社会言語学的分析のために，ランダムフォレストによる著者固有表現抽出を行なった．ランダムフォレストのテキスト分類への有効性，および著者固有表現抽出における有効

表 3　ジニ係数上位 10 語

中曽根	順位	語彙	ジニ係数	p 値	小泉	順位	語彙	ジニ係数	p 値
	1	ます. 助動詞.	0.538	0.017**		1	ある. 助動詞.	1.616	≪0.001*
	2	で. 助詞. 格助詞	0.505	0.001*		2	だ. 助動詞.	1.495	≪0.001*
	3	べし. 助動詞.	0.479	0.013**		3	べし. 助動詞.	1.043	≪0.001*
	4	う. 助動詞.	0.323	0.019**		4	です. 助動詞.	0.878	≪0.001*
	5	における. 助詞. 格助詞	0.283	0.035**		5	を. 助詞. 格助詞	0.620	≪0.001*
	6	の. 助詞. 格助詞	0.262	0.013**		6	たい. 助動詞.	0.608	≪0.001*
	7	など. 助詞. 副助詞	0.246	0.022**		7	が. 助詞. 接続助詞	0.585	≪0.001*
	8	の. 助詞. 連体化	0.240	0.017**		8	も. 助詞. 係助詞	0.573	≪0.001*
	9	に対して. 助詞. 格助詞	0.230	0.059		9	で. 助詞. 格助詞	0.553	0.001*
	10	ん. 助動詞.	0.202	0.156		10	ます. 助動詞.	0.484	≪0.001*

* 1% 信頼区間で有意.
** 5% 信頼区間で有意.

性を確認し，これを適用することで，著者固有の文体的特徴によって，著者固有の政治スタイルや個人的特徴の分析が可能となることを示した．同様の手法は，他の様々なテキストの社会言語学的分析にも有効であると考えられ，本研究は，潜在的に重要な文体論の応用分野であるされながら [4]，まだ実際の展開事例は少ない，コンピュータ社会言語学の領域を切り開くものである．今後，抽出された表現をより具体的に検討することで，それぞれの政治スタイルの分析を詳細に行なうとともに，著者固有の文体的特徴と，内容語の共起関係を検討することで，それぞれの政治家が強調するトピックや政策を明らかにする予定である．

参考文献

[1] AHRENS, K. People in the State of the Union: viewing social change through the eyes of presidents, *Proceedings of PACLIC 19, The* 19*th Asia-Pacific Conference on Language, Information and Computation* (2005), 43–50.

[2] ARGAMON, S., GOULAIN, J.-B., HORTON, R. and OLSEN, M. Discourse, power, and écriture féminine: text mining gender difference in 18th and 19th century French literature, *Abstracts of Digital Humanities* (2007), 161.

[3] ARGAMON, S., HORTON, R., OLSEN, M. and STEIN, S. S. Gender, race, and nationality in Black drama, 1850-2000: mining differences in language use in authors and their characters, *Abstracts of Digital Humanities* (2007), 149.

[4] ARGAMON, S., WHITELAW, C., CHASE, P., RAJ HOTA, S., GARG, N. and LEVITAN, S. Stylistic text classification using functional lexical features, *Journal of the American Society for Information Science and Technology*, **58**, 6 (2007), 802–822.

[5] BREIMAN, L. Random forests, *Machine Learning*, **45**, 5-23 (2001).

[6] DÍAZ-URIARTE, R. and ANDRÉS, ALVAREZ DE S. Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, **7**, 3 (2006), www.biomedcentral.com/1471-2105/7/3.

[7] KENNY, A. *The Computation of Style: an Introduction to Statistics for Students of Literature and Humanities*, Pergamon Press, Oxford (1982).

[8] YU, B. and UNSWORTH, J. An evaluation of text classification methods for literary study, *Abstracts of Digital Humanities* (2007), 157.

[9] 東照二 歴代首相の言語力を診断する, 研究社, 東京 (2006).

[10] 金明哲 助詞の分布における書き手の特徴に関する計量分析, 社会情報, **11**, 2 (2002), 15–23.

[11] 金明哲, 村上征勝 集団学習法による文章の書き手の同定, じんもんこん 2006 人文科学とコンピュータシンポジウム論文集 (2006).

[12] 草野厚 歴代首相の経済政策全データ, 角川書店, 東京 (2005).

[13] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸日本語形態素解析システム『茶筅』ver.2.2.3, (http://chasen.naist.jp) (2003).

[14] 大塚裕子, 乾孝司, 奥村学 意見分析エンジン: 計量言語学と社会学の接点, コロナ社, 東京 (2007).

[15] 鈴木崇史, 影浦峡 時代による総理大臣演説の文体的変化, じんもんこん 2006 人文科学とコンピュータシンポジウム論文集 (2006).

[16] 高瀬淳一 武器としての<言葉政治>－不利益分配時代の政治手法－, 講談社, 東京 (2005).

[17] 五百旗部真（編） 戦後日本外交史, 有斐閣, 東京 (2006).