

アメリカ合衆国大統領演説の計量テキスト分析

青木 宣康[†]河村 俊太郎^{††}鈴木 崇史^{†††}東洋大学社会学部[†]東京大学大学院教育学研究科^{††}東洋大学社会学部^{†††}

1. はじめに

演説は従来から政治指導者にとって、強力な武器として使用されてきた。演説力があるかどうかで、国民からの支持が得られるかどうかが決まると言っても過言ではない（川上，2009）。

とりわけ、アメリカ大統領は、世界に影響を与える者として、重要な役割を担っており、従来から、それらの演説は、言語学的、政治学的に重要な研究対象となってきた（ex., Ahren, 2005）。

一方、自然言語処理技術の進展とデータベースの発達によって、多様なテキストから多様な情報を抽出することが、欠かせない研究課題となっている。とりわけ、blog や Twitter をはじめとする断片的なテキストが多く産出され、また、domain adaptation の問題が自然言語処理分野で重要な課題になっている現在（Hara et al., 2010）、様々なジャンルのテキストの特徴と、そのテキストへの影響要因を検討することは、言語処理研究の観点からも重要な意義をもつ。

このような背景のもと、本研究では、過去 44 期 57 本のアメリカ大統領の演説を分析する。本研究では、予備的分析として、アメリカ大統領演説のテキスト特徴量への影響要因を探索的に探ることとし、単純な bag of words モデルと教師なし、教師あり学習を適用する。これによって、今後、アメリカ大統領演説から言語学的、政治学的含意を得るために有用な基礎的知見を得ることを目指す。

2. データ

1789 年から 2009 年までのアメリカ合衆国大統領演説、44 期 57 本を対象とする。Miller Center¹を参照し、演説のテキストファイルを作成、タイトル、スペースを削除した後、Tree Tagger²を

適用、すべての語の相対頻度を計算し、テキスト - 特徴量行列を作成する（bag-of-words モデル）。なお、活用形の統一は行わない。

3. 分析手法

3.1. カーネル主成分分析

まず、テキスト - 特徴量行列に対してカーネル主成分分析を適用する。これによって、テキストに影響を与える主要因を検討し、また、テキスト間の距離を可視化する。パラメータには $\sigma=0.2$ を与える。

3.2. ランダムフォレスト機械学習法

次に、主要因と考えられる時代、政党を教師信号として、ランダムフォレスト機械学習法（Breiman, 2001）を適用する。これによって、これらの要因のテキスト特徴量への影響を検討する。時代区分に関しては、政治的な背景も考慮し、2 クラス、5 クラスの 2 種類をラベルとして利用する。より具体的に、教師信号は、以下の通りである。

実験 1（時代 1）：第一次世界大戦終結前後（1789-1918, 1918-現在）,

実験 2（時代 2）：独立-南北戦争終結-第一次世界大戦終結-第二次世界大戦終結-冷戦終結-現在（1989-1865, 1865-1918, 1918-1945, 1945-1989, 1989-現在）,

実験 3（政党）：民主党（Democratic Party）、共和党（Republican Party）、その他,

実験の評価には、精度、再現度、 $F1$ 値のマイクロ平均を使用する。

4. 結果と考察

4.1. カーネル主成分分析の結果

図 1 はカーネル主成分分析の結果を示したものである。散布図中、大統領の略号でテキストの位置を示してある。中央よりやや右側に冷戦後の大統領が位置し、それ以前の大統領は中央より左側に位置する傾向があることから、第一主成分は時代をあらわすと解釈できる。これに対し、第二主成分の解釈は、自明ではない。

Quantitative text analyses of U.S. presidential speeches

[†] Noriyasu Aoki, Faculty of Sociology, Toyo University

^{††} Shuntaro, Kawamura, Graduate School of Education, University of Tokyo

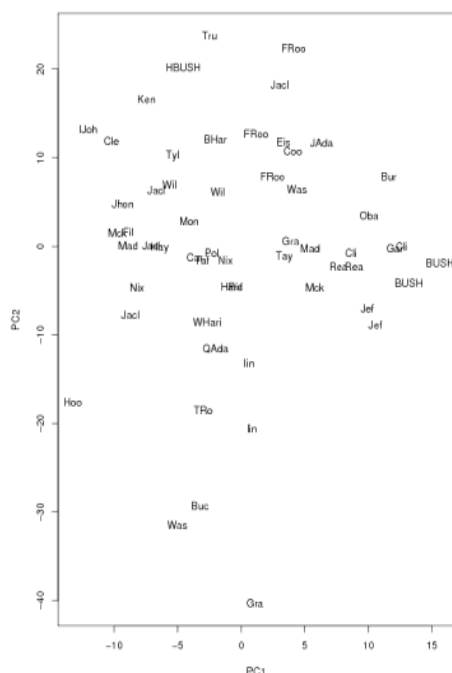
^{†††} Takafumi Suzuki, Faculty of Sociology, Toyo University

¹ millercenter.org

² www.ims.uni-

stuttgart.de/projekte/complex/TreeTagger

図1 カーネル主成分分析の結果



4.2. ランダムフォレスト機械学習法の結果

表1はランダムフォレスト機械学習法の結果を示したものである。2クラスの時代区分による分類実験（実験1）では、高いF1値が得られている一方、5クラスの時代区分による分類実験（実験2）では高いF1値が得られていない。また、政党別の分類実験（実験3）は、両者の中間程度の分類性能である。³

以上の結果から、第一次世界大戦終結前後では、アメリカ大統領演説に比較的大きなテキスト上の差異がある一方、より細かな時代区分、あるいは政党別では、相対的に小さなテキスト上の差異しかないことが指摘できる。

表1 ランダムフォレスト機械学習法の結果

クラス数		精度	再現度	F1値
1	2	95.24	97.37	96.15
2	5	32.50	-	-
3	3	46.98	43.81	44.76

5. おわりに

本研究では、57本のアメリカ大統領演説を用

い、bag of words とカーネル主成分分析、ランダムフォレストを利用して、これを分析した。

カーネル主成分分析の結果からは、時代が主要な影響要因であることが示され、さらに、ランダムフォレスト機械学習法の結果からは、そのような時代区分の影響のうち、第一次世界大戦終結前後に大きな時代区分の影響は観察されるものの、より細かな時代区分や政党の影響は観察されないことが示された。

本研究は、bag of words を用いることで、主要な影響要因の把握を目指した。今後、他の特徴量も利用し、また、細かく影響要因を区別することで、さらにアメリカ大統領演説のテキストの特徴を明らかにし、その言語学的、政治学的含意を探っていきたい。

謝辞

本研究は、科学研究補助金若手研究(B)「計算文体論による多種メディアテキスト解析（研究代表者：鈴木崇史、研究課題番号：23700288）」および、国立情報学研究所公募型共同研究「多種テキストからのコミュニケーション・スタイルの抽出ならびにその分析と応用（研究代表者：鈴木崇史）」より、一部支援を受けています。ここに記して謝意を表します。

文献

- Ahrens, K. (2005) People in the State of the Union: viewing social change through the eyes of presidents, *Proceeding of PACLIC19: the 19th Pacific Asia Conference on Language, Information and Computation*, 43-50.
- Breiman, L. (2001) Random forests, *Machine Learning*, 45(1), 5-32.
- Hara, T., Miyao, Y. and Tsujii, J. (2010) Evaluating the Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser, Bunt, H., Merlo, P., Nivre, J. (Eds.), *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing. Text, Speech and Language Technology*, 253-272, Springer, Dordrecht: Heidelberg: London: New York.

川上徹也 (2009) あの演説はなぜ人を動かしたのか, PHP 研究所, 東京.

³ 他のいくつかの時代区分、また、時代を限定して政党別の分類を試みたが、それぞれ、実験2、実験3と同様の傾向が得られた。