テキストマイニングによる
選挙候補者演説の分析

ソリューション本部　ソリューションサービス部

甲斐　尊賢

1. はじめに

インターネット上の様々なデータを分析し、自社の経営戦略等に役立てる手法のひとつとして、近年マイニングが注目されている。 なかでもテキストデータを入手し分析する環境が整備されてきており、手軽に利用できるようになってきている。本稿では、テキストデータの形態素解析エンジンであるMeCab とデータ解析環境であるR、(RMeCab)を用いて、2012年 9 月 26 日に実施された自由民主党総裁選挙、候補者5人の所見から候補者の分析を行ったので報告する。

2．解析エンジンと解析の対象

2. 1　解析エンジンについて

日本語の文章は、1文字ずつでは意味の無い語の連続であるが、単語として区切ることにより、初めて意味のある言葉になる。文章を読む際、人間は文章中の文字をつなぎあわせ、単語の連続として整理することを無意識の内に行うことができる。

しかしながら、コンピュータは日本語の文章を人間のように単語の連続として区切る(理解する)ことは出来ない。そのような理解をコンピュータ上でも可能とするために形態素解析と呼ばれる、自然言語を対象とした解析手法が出現した。形態素解析として、MeCab 、ChaSen、JUMAN、KAKASI などがある。この中で MeCab は京都大学等のプロジェクトで工藤拓氏により開発されたオープンソースの形態素解析エンジンであり、他の形態素解析処理と比較しても比較的精度の高い単語抽出が可能になっている。

本分析では、データ解析ツール R から MeCab の機能を呼び出すツール RMecsb を使用する。

表 1　MeCab の概要

項目	手法
解析モデル	bi-gramマルコフモデル
コスト推定	コーパスから学習
辞書引き アルゴリズム	Double Array
連接表の実装	2次元Table
品詞の改装	無制限多階層品詞
未知語処理	字種（動作定義を作成可能）

2. 2　自由民主党総裁選挙

2012 年 9 月 26 日に実施された自由民主党総裁選挙は表2の5氏が立候補し、当選者は次回衆議院総選挙後には首相に就任する確率が高いこともあり、熱を帯びた選挙戦が実施された。ここでは、立候補者の「総裁選候補者所見発表演説会」における所見発表演説をテキスト化したもの(自由民主党ホームページより)を対象とし、その内容から各候補者のポジションを分析した。表1に候補者の立候補演説の一部と選挙結果を示す。演説文の文字数は330文字から344文字でほぼ同程度の長さである。

表2　5人の候補者の演説内容

候補者	安倍晋三氏	石原伸晃氏	石破　茂氏	町村信孝氏	林　芳正氏
略称	安倍	石原	石破	町村	林
立候補演説文（途中で省略）	私は突然、首相の職を辞した。心からおわび申し上げる。責任をずっと考えたが、昨年の東日本大震災で３４万人が仮設住宅など困難な生活を強いられている。	谷垣総裁は自民、公明、民主の３党合意をより確かに実現するためにバトンを若い世代に託された。３年間の自民党の成果、路線をしっかりと守り、さらに時代を前に進めなければいけない。	議員、閣僚、首相であることは何かを成し遂げるための手段だ。私がなぜ今、総裁選に出馬するのか。それは今が祖国日本の国難だからだ。	力強い日本をつくり、次の世代にしっかりと渡していく信念に燃えて立候補した。民主党政権はいろいろな課題を解決できず、まさに国難だ。	日本の経済をなんとか再生させたい。再生をこの林芳正に任せてもらいたい、そういう強い決意で無謀とも思える総裁選に挑戦させていただいた。
文字数	334	343	332	330	335
１回目選挙	141	96	199	34	27
決選投票	108		89		

2. 分析手法と結果

2. 1　候補者所見からの形態素解析

5人の候補者所見テキストを RMeCab を用いて形態素（意味の最小単位）に分割し、名詞と形容詞の出現頻度を算出した。 各候補者毎で使用頻度の高い名詞を抽出した頻度表を表2に示す。

表2　名詞・形容詞の出現頻度表

	安倍	石原	石破	町村	林
日本	8	6	6	2	7
こと	6	3	4	9	5
ため	7	2	0	2	4
経済	5	0	3	4	3
的	3	0	2	3	6
憲法	4	4	5	0	0
今	3	0	5	0	2
人	3	4	0	0	3
保障	2	0	5	3	0
国	0	4	6	0	0
以下省略					

2. 2　候補者間の類似度（距離）

各候補者間の類似度を比較する手法として距離を求めた。候補者別の名詞・形容詞出現頻度を基に、各候補者間の距離を算出した。距離算出方法として代表的なユークリッド距離（表3）、およびキャンベラ距離（表4）を用いた。

用いる距離によりかなり異なった位置関係となる。

表３　各候補者間の距離（ユークリッド距離）

	安倍	石原	石破	町村
石原	23.4			
石破	21.2	20.0		
町村	22.4	23.6	20.5	
林	20.3	21.6	19.6	20.5

表４　各候補者間の距離（キャンベラ距離）

	安倍	石原	石破	町村
石原	108.7			
石破	103.9	109.0		
町村	101.6	112.1	103.3	
林	99.4	107.7	107.9	108.3

2．3　候補者のポジション分析

これらの距離マトリクスより、少数次元に配置する方法の一つである多次元尺度法により、2次元に配置表現を行った。キャンベラ距離による結果を図1に示す。また、距離マトリクスより階層型クラスター分析によりデンドログラムを作成した。(図2)順次近いものを統合してゆく過程をみることで、各候補者の関係を把握することが出来る。

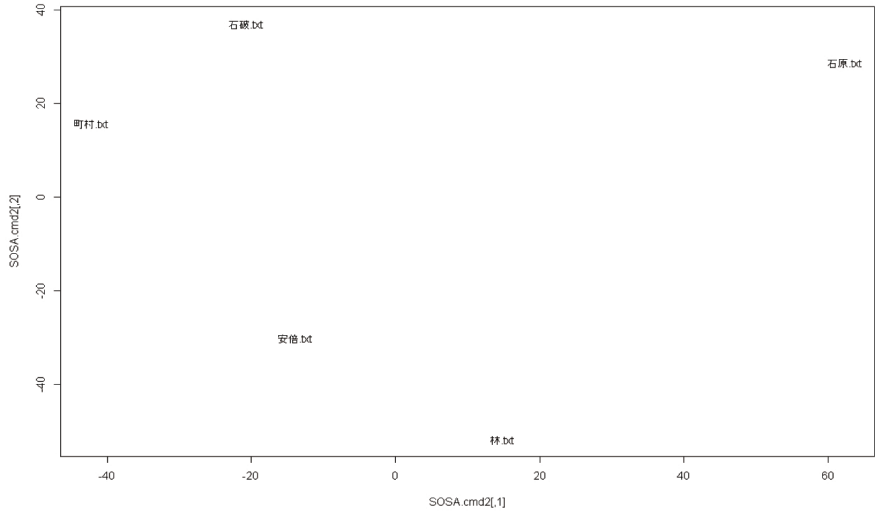

図1　キャンベラ距離による多次元尺度法結果

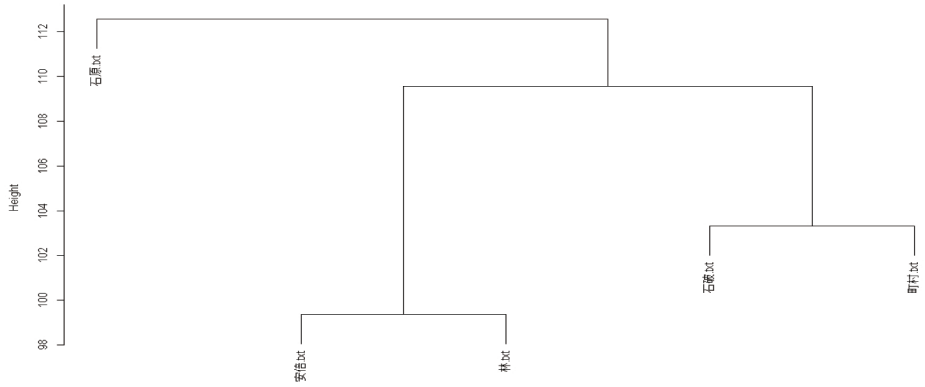

図2　デンドログラム(キャンベラ距離)

2．5　候補者とキーワードとの関連性

候補者と所信で使用されている言葉(名詞)との関連性をみるため、候補者別出現頻度3回以上の名詞の頻度表を算出した。

これをもとに、主成分分析を行い、第1主成分(寄与率 35%)、第2主成分(寄与率 21%)について、候補者の固有ベクトル、言葉のスコアをプロットして図3に示す。

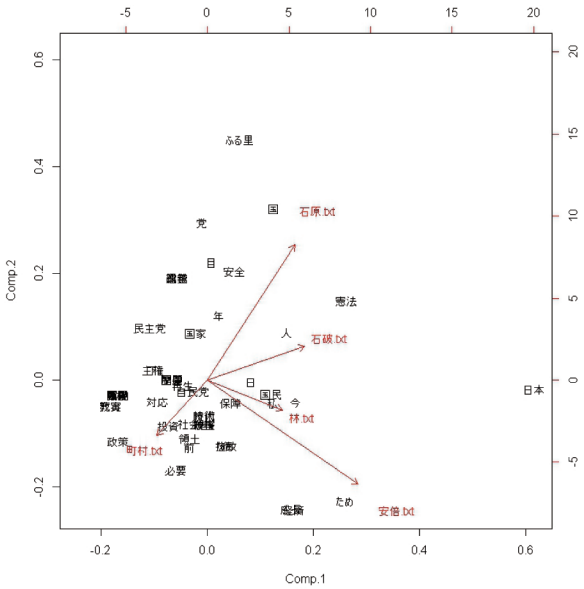

図3　主成分分析の結果

第1主成分は保守・リベラルを表す軸か、第2主成分は解釈が難しいがグローバル・国内重視を表す軸ともみられる。

これより、主要3候補（安倍、石破、石原）のうち2候補が決選投票に残った場合の決選投票の行方を図3より予想してみる（事前に予想した）。なお、決選投票では、より近い所見の候補に票が行くものとして考えた。

＜安倍Vs. 石破＞分割線の引き方によるが、安倍・町村・林対石破・石原となり接戦

＜安倍Vs. 石原＞石破の票の争奪となるが、安倍優勢

＜石原Vs. 石破＞石原対石破他となり石破圧勝

3. 結び

自由民主党の総裁選挙の候補者所見発表演説の文書の分析を行った。

分析結果は、選挙結果と比較しても、演説の文書からのみでもある程度の予測ができるという、興味深い結果となっている。また、テキストデータの解析環境である RMeCab は、使いやすく、適用可能な解析手法も豊富である。今後、このようなオープンソース解析環境のさらなる発展、整備に期待したい。

＜謝辞＞

本稿の執筆にあたり、多大なるご協力を頂いた株式会社空間情報の矢野公一氏に深く感謝します。

＜参考文献＞

1)　「テキストマイニング入門」（石田基広、森北出版株式会社、2008 年 12 月 22 日）
2)　「テキストデータの統計科学入門」（金明哲、株式会社岩波書店、2009 年 4 月28 日）