国会会議録に基づく短命大臣の特徴分析 第2報

大南勝 **掛谷英紀**

s1620760@u.tsukuba.ac.jp **kake@iit.tsukuba.ac.jp**

筑波大学

概要 本論文では、Web 上で公開されている国会会議録を言語資源に、機械学習によって、国務大臣としての資質に欠ける人物の特徴を見出すことを目的とする。そこで、長期的に大臣職を務め上げた人物(長期大臣)と、大臣就任後舌禍や不祥事によって辞任した人物(短期大臣)の分類を試みる。機械学習には最大エントロピー法を用いた学習と、判別分析法を用いた学習をそれぞれ実装し、その分類性能を比較する。機械学習により、短期大臣は国会の場にふさわしくない砕けた表現を多用するほか、高い理想や自身の頑張りを主張する発言が多い傾向がうかがえることが分かった。また、データサイズが大きく、同じ単語が大量にデータ中に含まれる場合や、特定の時代や政党の政策に関する単語など、クラス内で一部の人物のみが頻繁に使う単語を自動的にスクリーニングしたい場合には、判別分析法を用いた機械学習が有効であることが示された。

1 はじめに

近年、政治の世界において、政治的要職に就いた人物が問題発言や不祥事によって任期途中で辞任することが相次いでいる。平成 28 年 8 月 3 日から平成 29 年 8 月 3 日まで続いた第 3 次安倍第 2 次改造内閣では、今村雅弘(元復興大臣)、稲田朋美(元防衛大臣)らが立て続けに任期途中で辞任している。一方、内閣改造において重要ポストに再任され、その職務を長期にわたって務め上げる人物もいる。こうした人選を全てのポストに対して実施できれば、政治はより安定したものになる。

これまでも国会会議録を対象として自然言語処理を応用した研究はいくつかあるが[1-3]、これらは主に政治的主張の分類や類似度情報を提供するものであって、国会議員としての政治的資質の有無に関する情報は全く提供されない。筆者らは国会会議録を言語資源に、機械学習を用いて、長期的に大臣を務め上げる人物(長期大臣)と、大臣就任後舌禍や不祥事によって辞任した人物(短期大臣)の分類と、短命に終わる大臣の特徴を見出す研究を行っている[4]。本研究では、学習データや学習法に新たな改良を加えた結果を報告する。

2 手法

2.1 概要

本研究では、国会会議録に収められている国会答弁を言語資源として、教師あり学習によって文書を分類するシステムの構築を試みる。まず、国会会議録検索システム[5]から国会会議録検索システム検索用 API[6]を用いて国会会議録の文書データを入手する。

次に、取得した文書データを形態素解析ツール MeCab[7]を用いて形態素ごとに分割し、素性データを決定する。このようにして抽出した特徴語から訓練データおよびテストデータを作成する。その後、訓練データを用いて、各カテゴリ間の特徴を機械学習し、その学習したモデルにテストデータを入力して判定結果を得る。

機械学習には最大エントロピー法を用いた学習と判別分析法を用いた学習をそれぞれ実装し、分類性能の比較を行う。精度を算出する際には、共にクロスバリデーション(交差検証)を行う。本研究では最大エントロピー法を用いた機械学習のプログラムとして maxent[8]を使用する。以上の手順を図 1 に示す。

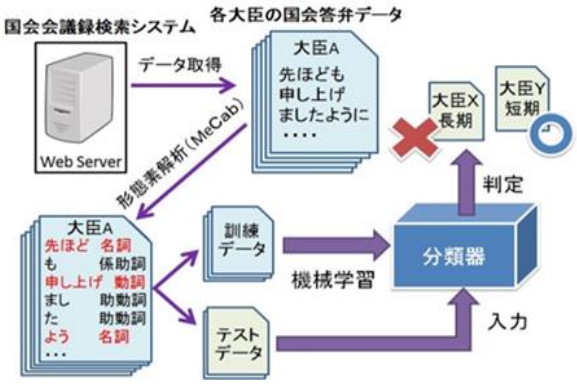

図 1 システムの概要

2.2 判別分析法

最大エントロピー法による学習では、各素性が文書データに出現するかしないかの 2 値情報しか用いない。単語の出現頻度情報を考慮した手法として、判別分析法を用いた機械学習を考える。本研究ではまず判別分析法を用いて、各素性について分離度

$$d = \frac{{\sigma_\beta}^2}{{\sigma_w}^2} \quad (1)$$

を求める。ここで ${\sigma_\beta}^2$ はクラス間分散、${\sigma_w}^2$ はクラス内分散を表しており、それぞれ

$${\sigma_\beta}^2 = (m_L - m_S)^2 \quad (2)$$

$${\sigma_w}^2 = \frac{n_L {\sigma_L}^2 + n_S {\sigma_S}^2}{n_L + n_S} \quad (3)$$

で与えられる。n_L , n_S はそれぞれ学習データの長期、短期大臣数である。また m_L , m_S は長期、短期カテゴリにおける素性の平均出現確率であり、${\sigma_L}^2$ と ${\sigma_S}^2$ は長期、短期カテゴリにおける素性の出現確率の分散である。クラス内分散を考慮することにより、クラス内で使用頻度に大きく差がある素性の影響を小さくすることができる。このように各素性に対して分離度 d を求め、分離度が閾値以上の素性を特徴語として抽出する。

次に、各文書における素性の出現確率は

$$\mathrm{P}(k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!} \quad (4)$$

で表されるポアソン分布に従うという仮定のもと各カテゴリで抽出した特徴語について確率モデルを求める。ここで λ は各カテゴリでの素性の出現確率平均、k はテストデータにおける素性の出現確率である。ただし、k は整数になるように、出現確率の小数点以下を切り上げて扱う。そして、特徴語ごとに算出した確率をナイーブベイズ的に掛け合わせることで、テストデータがそのカテゴリである確率が求められ、その確率が最大になるカテゴリにテストデータは属すると判定される。

3 大臣任期中の全発言を用いた分類

3.1 収集データ

本稿では長期大臣を「内閣改造で二度以上留任した人物」と定義する。一方、短期大臣を「内閣改造以外で任期途中に大臣職を辞任した人物」と定義する。収集対象は、昭和 62 年 11 月 6 日（第 1 次竹下内閣）から平成 28 年 8 月 3 日（第 3 次安倍第 1 次改造内閣）の期間内に内閣総理大臣を除く大臣職に就いていた人物する。上記の定義に該当する長期大臣は 14 名（延べ 15 名）、短期大臣は 48 名（延べ 49 名）となった。

次に、各カテゴリで発言の収集対象期間を次のように設定する。長期大臣については大臣就任日から退任日までとする。一方、短期大臣の収集対象期間は大臣就任日から不祥事発覚日前日までとする。これは、不祥事発覚後の国会での追及に対する答弁に現れる弁明などの特徴的な発言の影響を排除するためである。ただし、不祥事発覚日が不明な場合は退任日までとする。

以上の基準で発言の収集を試みたところ、長期大臣は上記の延べ 15 名全員の発言が取得できた。一方、短期大臣については在任期間が数日から数十日と極端に短い人物も含まれているため、10 名の発言は取得できず、計 39 名の発言のみ取得可能であった。

3.2 最大エントロピー法を用いた機械学習

本稿では、機械学習の際にカテゴリ間の発言数および役職の偏りが小さくなるように、発言数の少ない短期大臣を複数人組み合わせて長期大臣の発言数になるべく揃えてデータデータセットを作る。ここでは、長期・短期各 14 個のデータセットを作成する。

機械学習に用いた形態素は名詞・動詞・形容詞である。動詞・形容詞についてはそれぞれ文章中で現れた表出形から活用情報を省いた基本形を素性として使用する。分類結果を表 1 に示す。

全体の正解率は 75.0%となり、国会の発言をもとに長期大臣及び短期大臣を言い当てることがある程度可能であることが分かる。この判定においてα値の高い素性のうち特徴的な素性を長期カテゴリ・短期カテゴリそれぞれについて表 2 に示す。

表 1 分類結果（最大エントロピー法）

	再現率	適合率	総数	長期判定数	短期判定数
長期大臣	71.4%	76.9%	14	10	4
短期大臣	78.6%	73.3%	14	3	11
総数	75.0%	75.0%	28	14	15

表 2 上位素性（最大エントロピー法）

長期大臣		短期大臣	
東京電力	コントロール	明るい	究極
放射線	現実的	リゾート	役立つ
汚染	野田内閣	生きがい	完璧
原子力発電	無駄遣い	意図的	いっぱい
東日本	政権交代	人権	もう一歩
マネジメント	衆議院選挙	ゆとり	まじめ
実用	マニフェスト	あっち	未来永劫
シナリオ	一定の効果	憂慮	っぽい

短期カテゴリでは「あっち」、「いっぱい」、「っぽい」など国会答弁としてふさわしくない砕けた表現が多く見られる。また「究極の目標として」、「完璧を期す」、「未来永劫とは申しませんが」といった極端な表現を好む傾向がある。一方、長期カテゴリでは、「現実的に進める」、「実用化に向けた」など短期カテゴリとは対照的な素性が見られる。しかし、全体的に長期カテゴリの上位素性には東日本大震災や民主党に関する素性が多くみられる。このことから、発言に使われる表現の違いではなく、政党や発言の時期の違いを学習している可能性がある。これは、作成したデータセットにおいて、民主党政権時

代の大臣が長期カテゴリには延べ 10 名含まれているのに対して、短期カテゴリには 4 名しか含まれていないことが影響しているが考えられる。

3.3 判別分析法を用いた機械学習

本節では判別分析法を用いた機械学習を行う。データセットは 3.2 節で作成したものを使用する。

本実験では分離度に閾値を設け、分離度がその閾値以上の素性のみを使用して機械学習を行う。閾値を 0.3 から 1.3 まで、0.1 刻みで変化させた場合の正解率の推移を図 2 に示す。分離度の閾値が 1.0 の場合に正解率が 82.1%で最大となり、また閾値が 0.4 から 1.3 の範囲において最大エントロピー法による学習性能と同じか、それを上回る結果となった。最大正解率が得られた閾値が 1.0 の判定における、再現率、適合率をまとめたものを表 3 に示す。

また、この判定において長期カテゴリ・短期カテゴリそれぞれについて、分離度の高い素性のうち特徴的な素性を表 4 に示す。

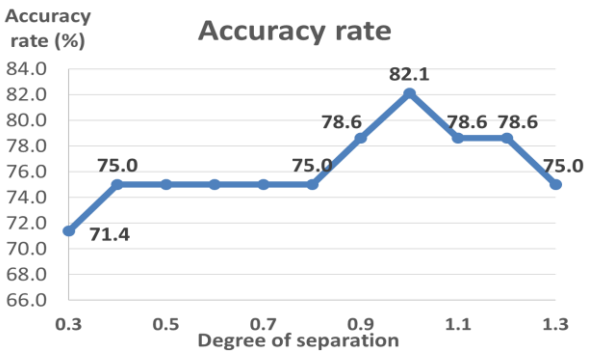

図 2 正解率推移

表 3 分類結果(判別分析法:閾値 1.0)

	再現率	適合率	総数	長期判定数	短期判定数
長期大臣	71.4%	90.9%	14	10	4
短期大臣	92.3%	76.5%	14	1	13
総数	82.1%	82.1%	28	11	17

表 4 上位素性(判別分析法)

長期大臣		短期大臣	
人材育成	ふう	見守る	もう一歩
被災地	真摯	とれる	信ずる
方向性	是非	ぶつかる	構え
丁寧	安定的	しよう	完璧
東日本大震災	シナリオ	生きがい	一生懸命
基本的	一定	もっともっと	すばらしい
一定の効果	僭越	くれる	喜ぶ
レビュー	現実的	悩ます	全力投球

表 4 において、長期大臣には「真摯に受け止める」、「まことに僭越ですが」など、相手を配慮した発言が目立っている。一方、短期大臣は「一生懸命頑張っている」、「全力投球をしている」といったように、自分の努力を一方的にアピールする発言が目立つことが分かる。さらに、短期カテゴリに見られる「しよう」という素性は「どうしようもない」を形態素解析した結果現れる素性であり、短期大臣はいざ厳しい質問・追及をされると開き直った発言をしてしまう傾向もうかがえる結果となった。また、長期カテゴリの上位素性を見ると、3.2 節の最大エントロピー法を用いた機械学習の際には上位に現れていた東日本大震災や民主党に関する素性の数は少なくなっている。このように、クラス内分散を考慮した判別分析法では、クラス内で特定の人物のみが頻繁に使う素性を自動的にスクリーニングし、学習から取り除くことが可能であることが示された。

4 大臣就任後 180 日以内の発言を用いた分類

3 章では大臣期間中の発言すべてを用いて、分類実験を行った。その際、長期大臣は就任直後の発言から、長く大臣職を務めた後の発言まで含まれている。一方、短期大臣はその任期の短さから、就任直後の発言が大半を占めている。そこで本節では、職務に慣れてきた後の答弁に現れる特徴的な発言の影響を小さくするため、長期・短期大臣ともに使用する発言を大臣就任後 180 日以内の発言に限定して分類を試みる。

4.1 データ収集

3 章で収集したデータは長期大臣が 15 人と少ない。そこで本章では、長期大臣の定義を「内閣改造で二度以上留任した人物」から「新内閣発足および内閣改造で二度以上留任した人物」に拡張する。短期大臣は 3 章同様「内閣改造以外で任期途中に大臣職を辞任した人物」と定義する。また収集対象に、第 3 次安倍第 2 次改造内閣で大臣職に就いていた人物を追加する。調査した結果、長期大臣は 44 名(延べ 48 名)、短期大臣は 50 名(延べ 51 名)が該当した。

4.2 最大エントロピー法を用いた機械学習

前章では発言数と役職の偏りが小さくなるよう工夫してデータセットを作成した。しかし、カテゴリ間で政党・および時期の偏りが大きくなり、機械学習の結果に影響していた。そこで本節では、発言数、政党、時期の偏りが小さくなるようデータセットを作成する。その際、短期カテゴリにおいて、大臣だけではデータが不足するため、短期副大臣のうち、宮路和明、平田耕一、原田義昭を加えてデータセットを作成する。作成した長期、短期それぞれ 25 個のデータセットを用いて機械学習を行う。

機械学習に用いた形態素は名詞・動詞・形容詞であ

る。動詞・形容詞についてはそれぞれ文章中で現れた表出形から活用情報を省いた基本形を素性として使用する。分類結果を表 5 に示す。

表 5　分類結果(最大エントロピー法)

	再現率	適合率	総数	長期判定数	短期判定数
長期大臣	80.0%	76.9%	25	20	5
短期大臣	76.0%	79.2%	25	6	19
総数	78.0%	78.0%	50	26	24

4.3　判別分析法を用いた機械学習

次に判別分析法を用いた分類を行う。本節ではより一般的に使われる言葉で分類するため、両カテゴリともに 10 回より多く出現する素性のみに対して、分離度を求め特徴語を決定する。使用素性は名詞・動詞・形容詞とする。本節では、分離度の閾値を 0.10 から 0.24 まで 0.01 ずつ変化させた場合の正解率推移を図 3 に示す。本実験での最大正解率は 74.0%となり、最大エントロピー法を用いた機械学習と比べて、4 ポイント低い結果となった。これは使用する発言を減らしたことで、頻度の偏りが小さくなったためと考えられる。閾値を 0.14 とした場合の再現率、適合率をまとめたものを表 6 に示す。

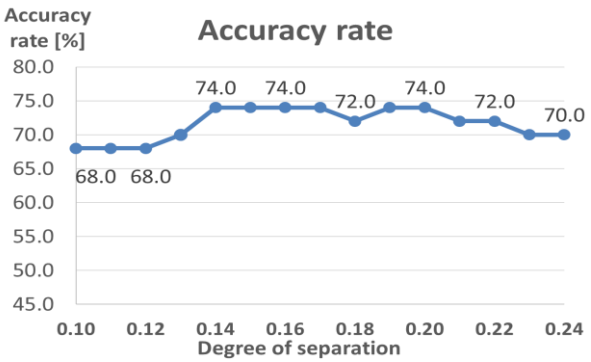

図 3　正解率推移

表 6　分類結果(判別分析法:閾値 0.14)

	再現率	適合率	総数	長期判定数	短期判定数
長期大臣	76.0%	73.1%	25	19	6
短期大臣	72.0%	75.0%	25	7	18
総数	74.0%	74.0%	50	26	24

また、この判定において分離度が 0.14 以上の素性のうち特徴的な素性を長期・短期それぞれについて表 7 に示す。長期大臣は政策に関する具体的な答弁がなされている傾向が見られる。一方で、短期大臣では現状の報告よりも、「思い」、「気持ち」、「考え」に言及することが多く、して「もらう」、して「くれる」のように責任逃れ、他人任せにする傾向が見られる。

表 7　上位素性(判別分析法)

長期大臣		短期大臣	
活用	実現	いずれ	求める
一つ	政策	守る	事情
重要	部分	報告	もらう
年度	数字	受ける	くれる
形	新しい	今度	思い
効果	例	私	お答え
必要	一部	次第	気持ち
目指す	基本的	聞く	考え

5　おわりに

本研究では、長期的に大臣を務める人物と、任期途中で辞任する人物の発言の違いを機械学習によって学習し、各カテゴリの特徴を抽出した。長期大臣は相手を取り込みながらうまく納得させるような発言が目立つほか、製作について現実的、具体的な答弁をする傾向がうかがえる。一方、短期大臣は国会の場にふさわしくない砕けた表現を多用する傾向があるほか、高い理想や自身の頑張りを主張する、厳しい質問をされた場合に責任逃れ、他人任せな姿勢を見せる傾向が見出せた。

分類実験では、最大エントロピー法と、単語の頻度情報を考慮した判別分析法を用いた場合の分類性能を比較した。同じ単語が大量にデータ中に含まれる場合や、クラス内で特定の人物のみが頻繁に使う単語を自動的にスクリーニングしたい場合には判別分析法を用いた機械学習が有効であることが示された。

本稿では、発言に基づく人物の信頼度評価が可能であることを確認した。将来、閣僚候補者の身体検査や有権者への投票支援、さらには採用活動や昇進などの人事活動支援への応用が考えられる。

参考文献

[1] 東、橋本、掛谷:Web 上の言語資源に基づく国会議員の分類, 言語処理学会第 17 回年次大会, 2011

[2] 東、掛谷:自己組織化マップによる国会議員のツイッター分類, 第 6 回メディア情報検証学術研究会, 2010

[3] 東、掛谷:国会議員のツイッター分類とその応用, 言語処理学会第 18 回年次大会, 2012

[4] 大南、掛谷: 国会会議録に基づく短命大臣の特徴分析, 言語処理学会第 23 回年次大会, 2017

[5] 国会会議録検索システム,http://kokkai.ndl.go.jp/

[6] 国会会議録検索システム検索用 API, http://kokkai.ndl.go.jp/api.html

[7] MeCab http://mecab.sourceforge.net/

[8] Masao Uchiyama. Maximum Entropy Modeling Package. http://www2.nict.go.jp/x/x161/members/mutiyama/software.html, 2006