N-001

# ランダムフォレストを用いた国会会議録のイデオロギー分析
# Analysis of the ideology in the Diet Record using Random Forests

中川 侑* 武田 拓也* 吉元 涼介** 芳鐘 冬樹***
Atsumu Nakagawa Takuya Takeda Ryosuke Yoshimoto Fuyuki Yoshikane

## 1. 研究背景

国会会議録は, 国会における議事内容や発言などの審議の様子を記録した公文書であり, 現在は国会会議録検索システム[1]により, Web で全文が公開されている. 国会会議録検索システムは国立国会図書館が提供するデータベースで, 1947 年の第 1 回国会から現在にいたるまでの衆参両院の本会議, 委員会の議事録を対象とした全文検索をはじめ, 議事内容や日付, 発言者などを指定した検索が可能である[2].

国会会議録を対象とした研究としては, 政治学, 社会学的なアプローチによる分析[3, 4, 5, 6, 7]が行われている一方で, 国会会議録を日本語の話し言葉を記録した大規模コーパスとして用い, オノマトペの使用傾向[8]や, 品詞の共起頻度[9]を明らかにする研究など, 言語学的なアプローチによる分析も行われている.また, 政治学の領域でテキストを分析する研究には, 精読による質的な内容分析[10, 11]がある一方で, 特定の内容語に着目して政策や政治的志向を明らかにする分析や, 機能表現に着目した文体分析[12], 語彙の多様性や名詞出現頻度などの特徴量をもとにした分析[13]において計量的手法が適用されている.

情報検索や機械学習の分野では, 文書をあらかじめ与えられたカテゴリに分類するテキスト分類の研究[14]が行われてきたが, 政治領域のテキストをイデオロギーに基づき分類する研究は, ごく僅かに見られる程度である[13]. テキスト中からイデオロギーを客観的に計る指標を得ることは困難であることが指摘されている[15].

## 2. 研究目的

本研究は, 国会会議録のテキストの計量的分析を通じて, テキストそのものの言語的な内容の吟味ではなく, 発言の計量的な特徴を比較することを目的とする. 具体的には, 国会会議録に収録されている議員の発言を機械学習によって分類することを通じ, 分類に有効なテキストの特徴量に基づいた, 政党（イデオロギー）の違いによる傾向を分析する.

## 3. 分析対象・手法

### 3.1 分析対象

国会会議録検索システムで閲覧できる第 183, 186, 189 回通常会衆議院予算委員会及び分科会の会議録から, 以下の通りにテキストを抽出した. ①公述人, 参考人は議員でないため, 発言を除外した. ②内閣総理大臣, 政務三役（国務大臣, 副大臣, 大臣政務官）を除く議員の発言を抽出する. 総理大臣をはじめとする閣僚, 政務三役の答弁は, 担当する行政庁の影響を受けるため, 発言を除外した. ただし, 先述の役職者であっても, 委員として行う質問については分析データに含めた. ③本研究では, 各議員のイデオロギーの類似や相違が, その所属政党に表れると仮定し, イデオロギーの区分には衆議院に議席を有する国政政党を用いた. そのため, 質疑のはじめに議員が名乗る政党名を, その議員の所属政党としてデータに付与した. なお, 当該期間中に所属政党に変更があった場合でも, この期間内ではじめに名乗った政党名を所属政党として付与している.

分析対象としたデータの概要を表 1 に示す. なお議員数は, 委員, 分科員の肩書の違いを考慮せず数える. また, この期間内には, みんなの党, 日本維新の会, 結いの党, 維新の党などの第三極政党が新たに結成され, 短い期間で分裂, 吸収, 合併がなされた. そのため本研究では, これらの政党をまとめて「維新の党等」として扱う.

表 1　分析対象の概要

| 政党 | 発言議員数 |
|---|---|
| 自民党 | 176 |
| 維新の党等 | 132 |
| 民主党 | 97 |
| 公明党 | 46 |
| 日本共産党 | 38 |
| 生活の党 | 10 |
| 計 | 499 |

### 3.2 分析手法

本研究の分析には, 集団学習アルゴリズムの一手法であるランダムフォレストを用いる. ランダムフォレストは高い分類精度が報告されており, また使用した説明変数の分類への影響度を算出することが可能である. さらに, 多変量解析で問題となる, 変数間の多重共線性の影響を回避することが可能である. 以下, 分析手法の詳細について述べる.

まず, 対象テキストに対し, MeCab を用いて形態素解析を施す. 次に, 形態素解析の結果をもとに議員ごとの特徴量ベクトルを作成し, 所属政党をそのクラスとして付与する. その特徴量ベクトルをもとにランダムフォレストによる分類を行い, その精度や, 寄与の大きな特徴量を明らかにする.

テキストの特徴量として, 以下に示す 14 の要約統計量を用いる. それぞれ, 形態素を単位とする最大文長, 最小文長, 平均文長, TTR（Type-Token Ratio：延べ語数-異なり語数比）, Simpson の多様度指数 *D*, 及び, 名詞, 動詞, 形容詞, 副詞, 助詞, 接続詞, 助動詞, 連体詞, 感動詞の各品詞が全形態素数に占める割合である.

* 筑波大学情報学群知識情報・図書館学類 College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba
** 筑波大学大学院図書館情報メディア研究科 Graduate School of Library, Information and Media Studies, University of Tsukuba
*** 筑波大学図書館情報メディア系 Faculty of Library, Information and Media Science, University of Tsukuba

## 4. 結果

ランダムフォレストによる分類結果を表 2 に示す. 行と列は, それぞれ, 議員の所属政党と分類先の政党であり, 値は該当する議員数である. 所属政党ごとの分類の精度を最右列に示す. 精度の高さは, 所属する議員の発言が特徴的である（他の政党との差異が大きい）ことを示唆する.

表 2　政党ごとの分類結果

| | 自 | 維 | 民 | 公 | 共 | 生 | 精度 (%) |
|---|---|---|---|---|---|---|---|
| 自民党 | 124 | 28 | 20 | 1 | 3 | 0 | 70.5 |
| 維新の党等 | 42 | 57 | 28 | 1 | 4 | 0 | 43.2 |
| 民主党 | 23 | 27 | 45 | 1 | 1 | 0 | 46.4 |
| 公明党 | 30 | 11 | 1 | 2 | 2 | 0 | 4.3 |
| 共産党 | 4 | 5 | 4 | 0 | 25 | 0 | 65.8 |
| 生活の党 | 2 | 3 | 3 | 2 | 0 | 0 | 0.0 |

重要度（分類への寄与）の高い特徴量について, 上位 5 つを表 3 に示す. これらのうち, 上位 2 つである「感動詞比率」と「TTR」について, 政党ごとの平均値を算出した. その散布図を図 1 に示す.

表 3　分類における特徴量の重要度

| 特徴量 | 重要度 |
|---|---|
| 感動詞比率 | 48.6 |
| TTR | 35.9 |
| 接続詞比率 | 32.5 |
| 副詞比率 | 29.2 |
| Simpson の多様度指数 $D$ | 26.9 |

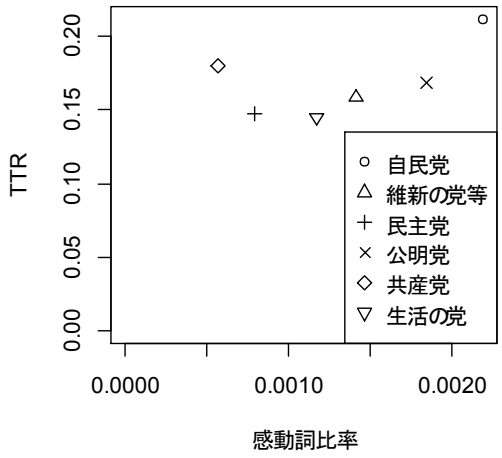


図 1　政党ごとの散布図

## 5. 考察

自民党と共産党の分類精度が比較的高い（70.5%と 65.8%）ことから, これらの政党の発言は独自性が高いことが示唆される. 維新の党等と民主党では, それぞれを混同する誤りが比較的多かった（28 人と 27 人）. 民主党は自民党の, 維新の党等は自民党や民主党の流れを汲んでいることから, これらの党を混同する誤りが多かったと考えられる. 公明党は 2016 年現在, 自民党と連立を組む与党であるが, この分類においても自民党との混同が多かった（30 人）.

感動詞比率が上位の自民党と公明党に着目すると, TTR が高く, 感動詞を含めた語彙が多様である. 感動詞比率が下位である共産党と生活の党は, 表 1 に見られるように, 議席数の少ない小規模会派であり, 限られた持ち時間で効率的に議論を進めるために, 感動詞を減らし, 議論の本題の部分に重きを置いていると考えられる.

## 6. 結論

本研究では, イデオロギーの区分として政党を用い, 国会における議員の発言について, 政党を特徴付ける要素の検討を行った. ランダムフォレストによる議員の発言の分析の結果, 感動詞の比率と語彙の多様性（TTR）について政党間の差異が大きい傾向が示唆された. 今後は, イデオロギーそのものの把握に直接つながるような特徴量を検討し, 分析に用いることで, 議員のイデオロギーが国会での発言にどのように反映されるのかを明らかにしていきたい.

### 参考文献

[1] 国立国会図書館. 国会会議録検索システム. http://kokkai.ndl.go.jp/.
[2] 大山英久. 国会会議録フルテキストデータベース. 情報知識学会誌. 1997, 7(1), pp. 19-26.
[3] 尾崎 正宗, 掛谷 英紀. 国会会議録の主張文取り出しおよびその要約. 言語処理学会第 20 回年次大会発表論文集, 2014, pp. 504-507.
[4] 丸山和昭. 「カウンセリング」のポリティクス：国会議事録の計量テキスト分析を中心に. 日本教育社会学会大会発表要旨集録. 2007, (59), pp. 331-332.
[5] 橋本 鉱市. 戦後高等教育政策におけるイシューとアクター--国会・文教委員会会議録の計量テキスト分析. 東北大学大学院教育学研究科研究年報. 2007, 56(1), pp. 71-87.
[6] 藤末健三. 自由貿易協定に関する民主党国会議員発言の政権交代前後の変化 -データマイニング手法を用いた国会議事録の分析-. アジア太平洋研究科論集. 2011, (22), pp. 1-20
[7] 澤勢一史, 延原肇. 形式概念分析を用いた国会議事録の束構造可視化と時空間解析. 電子情報通信学会技術研究報告. 2013, 112(465), pp. 17-19.
[8] 平田 佐智子, 中村 聡史, 小松 孝徳, 秋田 喜美. 国会会議録コーパスを用いたオノマトペ使用の地域比較. 人工知能学会論文誌. 2015, 30(1), pp. 274-281.
[9] 服部匡. 名詞と尺度的形容詞類の共起傾向の推移：国会会議録のデータから. 同志社女子大学学術研究年報. 2011, (62), pp. 113-141.
[10] 坂井誠. オバマ政権下の諸政策に関する政治経済的分析 1：政策思想と就任 1 年目の初期政策. 恵泉女学園大学紀要. 2010, 22, pp. 65-92.
[11] 笘米地真理. 尖閣諸島をめぐる「領有権問題」否定の起源：政策的解決への可能性. 公共政策志林 = Public policy and social governance. 2015, (3), pp. 139-153.
[12] 鈴木崇史, 影浦峡. 総理大臣国会演説における基本的文体特徴量の探索的分析. 計量国語学. 2008, 26(4), pp. 113-122..
[13] 鈴木崇史, 影浦峡. 総理大臣演説における語彙多様性の変化. 日本行動計量学会大会発表論文抄録集. 2007, 35, pp. 273-276.
[14] 湯浅夏樹, 上田徹, 外川文雄. 大量文書データ中の単語間共起を利用した文書分類. 情報処理学会論文誌. 1995, 36(8), pp. 1819-1827.
[15] 畑中允宏, 村田真樹, 掛谷英紀. 新聞社説・国会議事録に基づく言論のイデオロギー別分類. 言語処理学会第 15 回年次大会発表論文集. 2009, pp. 408-411.