政治的テキストの文法
—機械学習のための政治的テキストデータの構造—*

重村壮平 † 宋財泫（ソンジェヒョン） ‡

2017 年 5 月 24 日

概要

本稿の目的は、政治的テキストを高精度かつ機械的に分類する上で最適なテキストデータの構造を提示することである。これまで政治的テキストは、政党や政治家の政策位置の推定など、様々な場面で用いられてきた。これらの分析では、テキストを「ラベル付け」し「分類」する必要があり、多大なマンパワーを必要とする。しかし近年、機械学習アルゴリズムとこれを支えるハードウェアの進歩により、テキストデータを効率的かつ機械的に分類することが可能になった。本稿では、機械的処理に適したテキストの「構造」を提案するために、構造主義の知見を援用し、選挙公約を交換可能 (exchangeable) な意味を有する最小単位の集合として捉え、これを「公約素」と呼称する。この公約素を機械学習の応答変数として、公約素が選挙公約に含まれるか否かを機械的に予測し、本稿が提示したテキスト構造の有効性を検証した。検証の結果、本稿のテキスト構造は、既存のテキスト構造に比べて高いパフォーマンスの予測精度をもたらすことが明らかになった。すなわち、選挙公約を公約素で表現することによって、政治的テキストに含まれる情報量の損失を最小にし、精度の高い予測を行うことが可能であるとの結果が得られた。

---

* 日本選挙学会 2017 年度研究会報告論文 分科会 H（ポスターセッション）「選挙研究のフロンティア」

† 神戸大学法学研究科博士課程後期課程。E-mail: s.sohei@icloud.com

‡ 神戸大学法学研究科博士課程後期課程。日本学術振興会特別研究員 (DC)。
E-mail: tintstyle@gmail.com, URL: http://www.jaysong.net

1　はじめに

近年、ビックデータに対する関心が高まっている。その背景には、インターネットの普及と、それに伴うデータの入手コストの低下がある。誰もがビッグデータにアクセスできるようになった現在、政治学においてもビッグデータに対する関心は高く[1)]、「議会の議事録」「政党の政権公約」「政治家の演説文」などの政治的テキストデータを用いた研究は増加の傾向にある。ビックデータは、豊富な情報を含み、それゆえに因果推論に有用であるとの指摘があり、広く活用されることが望まれている (Grimmer, 2015)。それにもかかわらず、テキストデータを適切かつ容易に分析する方法は開発の途上にあり、政治的テキストデータの活用は望み通りに進んでいない (Monroe and Schrodt, 2008)。

テキストデータの主な分析手法は、内容分析 (content analysis) である。内容分析では、人間がテキストを一つ一つ精読し、適切なカテゴリにコーディングする必要がある。政治的テキストデータは情報量が多いため、適切に分析することで、客観的な見地から、様々な「政治的現象」を説明できる。しかし、この利点を裏返せば、分析に用いるためには、政治的テキストが包含する莫大な情報量を的確に処理する必要があり、多大なコストを負担しなければならない。供給される情報が幾何級数的に増加すると、一人の人間が精読できる分量を大幅に超え、分析は困難を極める。つまり、データの分類におけるヒューマンパワーの必要性が、テキスト分析を実施する上で障壁となり、政治的テキスト分析の興隆を妨げてきた (Hillard et al., 2008)。しかしながら、統計的・機械的学習手法の発展、ならびにハードウェアの進歩は、ビックデータの内容分析にかかるコストを大幅に低下させうる[2)]。そこで本稿は、機械学習に最適な選挙公約データの構造を提案することで、テキスト分析の障壁を取り除くことを目的とする。

公約は、政党の「政権公約」と候補者の「選挙公約」に分類できる。選挙公約は、政権公約に比べて、量的な面、質的な面の双方において優れている。一国内において候補者を擁立する政党の数は十数個に止まるものの、候補者の数は数百名に上るため、選挙公約は量的に充実している。さらに、候補者は選挙公約を通して実現したい政策を有権者に提示するため、選挙公約は候補者の選好を如実に反映するデータとして位置付けられる。つまり、選挙公約データを適切に分析できる方法を提示することで、候補者の政策位置を客観的かつ相対的に比較できるようになる。

[1)] Political Science & Politics 第 48 号におけるビックデータに関する論叢からも、政治学におけるビックデータへの関心の高まりを確認できる。

[2)] 政治学におけるテキスト分析の潮流ついては、Wilkerson and Casas (Forthcoming) が詳しい。

選挙公約から政策位置を統計的に推定するためには、公約を項目ごとにラベル付けする必要がある (品田, 2006)。テキスト分析に用いられる自然言語処理では「情報の抽出・構造化」と「同意・含意関係の認識[3)]」が主たる問題であり、さらに、これらの問題は「ラベル付け」問題に分解できる (乾・浅原, 2006)。選挙公約を用いた分析においても、その根幹には「ラベル付け」問題があり、この問題の解決なしに、機械的にテキストを分類することは難しい。選挙公約を効率的にラベル付けできるデータの構造と分析手法の提示は、選挙公約分析の発展において必要不可欠である。

本稿は以下の順に沿って議論を進める。第 2 節では、政治的テキストデータ分析の変遷と、政治的テキスト、とりわけ選挙公約データの構造に関するこれまでの研究をレビューする。第 3 節では、テキストデータのコーディング方法について紹介する。具体的には、第 3.1 小節は「人力」によるコーディング、第 3.2 小節は「機械」によるコーディングを紹介する。第 4 節では、機械的処理に適した選挙公約データの構造を紹介する。第 5 節では、機械学習による選挙公約の分類について、「データ（第 5.1 小節）」「データの前処理（第 5.2 小節）」「モデル（第 5.3 小節）」の順に詳述する。第 6 節では、実際に機械学習を用いて選挙公約を分類し、その結果を評価する。最終第 7 節において本研究の結論を述べる。

2　先行研究

政治的テキストデータは、たとえば、国家元首の言説を「肯定的」「否定的」「好戦的」等に分類 する研究 (e.g., Volkens et al., 2014)、イデオロギー空間上に政治家のポジションをスケーリングし、政治家の政策的立場を推定する研究 (e.g., Budge et al., 2001; 猪口, 1983; Klingemann et al., 2006; 小林・堤, 2000a,b,c; Laver et al., 2003; 品田, 2002, 2010; Slapin and Proksch, 2008) などに用いられる。これらの研究例から、テキスト分析の主たる目的は、「分類 (classification)」と「スケーリング (scaling)」にあるといえる (Grimmer and Stewart, 2013)。本研究では、テキストデータの「分類」に焦点を当てる。その理由は、テキストを適切なカテゴリに分類できなければ、因子分析や主成分分析を用いて行うスケーリングの妥当性を担保できないからである。

では、「分類」と「スケーリーング」という目的が確立されている政治的テキスト分析において、どのようなデータが用いられているのだろうか。たとえば、議会における議

[3)] 乾・浅原 (2006) は、「同意・包含関係の認識」を「抽出された情報どうしの意味的な類似性・包含関係を認識する処理」と定義する。

員の発言 (e.g., Lauderdale and Herzog, 2016; Proksch and Slapin, 2010; Quinn et al., 2010)、ブログや SNS 上における有権者や議員の投稿記事 (e.g., Beauchamp, 2017; Hopkins and King, 2010; O'Connor et al., 2010; 上ノ原, 2014)、マスコミによる報道記事 (e.g., Althaus et al., 2001; Boomgaarden and Vliegenthart, 2007)、選挙公約 (e.g., Laver et al., 2003; Lowe and Benoit, 2013) などが分析に用いられており、テキストデータの種類は多岐に及ぶ。本研究で用いるデータは、選挙公約である。選挙公約は、議会・委員会などの議事録に比べて、多くのアクターの情報を内包する。また、選挙期間中の情報の収集に際し、有権者が接触・参照する媒体の一つでもある[4)]。選挙公約は、データの量が豊富であり、質も担保されているため、分析に値する。そのため、選挙公約データを容易に取り扱える分析手法の開発は喫緊の課題といえる。

選挙公約を用いた研究は、日本と欧米を中心に発展してきた。選挙公約データの適用範囲は、政党の応答性 (e.g., 大村, 2012)、争点投票 (e.g., 小林, 1997, 2008; 小林他, 2014; 堤, 1998)、政策（政党間）競争 (e.g., Catalinac, 2016; 品田, 2002, 2010) など、政治の様々な側面に及び、これらの研究は政治学における理論の発展に貢献してきた。しかし、これらの研究で用いられるデータの生成過程に関心を寄せる研究は非常に少ない。ほとんどの研究が共通のデータに依拠しているため、データ生成過程の妥当性について議論の余地が生まれにくかった。例外的にデータ生成過程に言及しているのは、「マニフェスト国際比較プロジェクト（Comparative Manifesto Project； 以下、CMP と略す)[5)]」のデータを用いた研究 (e.g., Benoit et al., 2009; 大村, 2012)、品田 (1998, 2006) のデータを用いた研究 (e.g., 梶原, 2014) である。ここで強調すべきは、これらの研究が「公約分析には多くのコストを要する」との見解を強調する点である。このようにコーディングに要する時間的コストの負担が大きく、それゆえに研究者自身で独自のデータセットを構築することは困難を極める。テキストデータが持つ豊富な情報量にもかかわらず、テキストデータの利用を試みる研究者が数少ないの理由はコストに起因する場合が多い。

なぜ、テキストの「分類」には大きなコストがかかるのだろうか。その背景には、政治的テキストの「分類」を目的とする多くの研究が、ヒューマン・コーディング依拠している (Laver et al., 2003) という実態がある。コーダーは、テキストを注意深く読み込み、予め用意したカテゴリ番号を付けて、テキストを分類しなければならない。選挙公約のように数万から数十万の行で構成されるデータを人力でコーディングする場合、作業時間が数ヶ月に及ぶことも珍しくない。作業時間を短縮するために「Amazon Mechanical

[4)] 選挙公約は、選挙公報に記載されており、多くの有権者が接触すると考えられる。なお、選挙公報の発行は、各都道府県の選挙管理員会による発行が義務付けられている（公職選挙法第 167 条)。

[5)] CMP の詳細については Budge et al. (2001); Klingemann et al. (2006) を参照されたい。

Turk (以下、AMT と略す)」や「Yahoo!クラウドソーシング」のサービスを利用し、仕事を分散させることは可能だ (Benoit et al., 2016)。しかし、コーダーには高い熟練度が要求されるため (Volkens, 1992)、コーディングに精通しない人への業務委託は難しい。

しかし近年、統計モデリングの精緻化、ならびにデータを解析するハードウェアの進歩に伴い、コーディングに要する時間を大幅に短縮する手法が提唱されている (Sebastiani, 2002)。その代表例が、機械学習と自然言語処理である (Wilkerson and Casas, Forthcoming)。政治学においても、ヒューマン・コーディングを代替する方法として、テキストデータの機械的分析に対する関心は高く (Grimmer and Stewart, 2013)、実際に政治テキストデータの機械的分類が十分に実用化可能であることを示した研究は複数ある (e.g., Hopkins and King, 2010; Quinn et al., 2010; 上神・佐藤, 2009)。

本稿の目的は、これまでの研究を土台として、機械学習の性能を向上させる選挙公約データの構造を提案することである。そこで、先行研究の選挙公約データの構造を明らかにし、その構造の利点と改善点を確認したい。ここでは、データの収集から作成に至るまでの過程を詳細に紹介する品田 (1998, 2006) に基づき議論を進める。

品田 (1998, 2006) の選挙公約データは、「対象」「内容」「賛否」の三次元構造である。たとえば、選挙公約データにおける i 番目の候補者の n 個目の選挙公約を、

$$\begin{aligned} \text{公約}_{i,n} = 41\text{r}0\text{w} = \{&\textbf{対象} = \text{国民 (41)},\\ &\textbf{内容} = \text{政治改革 (r0)},\\ &\textbf{賛否} = \text{改革 (w)}\} \end{aligned}$$

と表現する。各要素がとりうる値は、対象が 34、内容が 117[6]、賛否が 3 個[7]である。対象と内容が取りうる値は非常に多く、ほとんどの選挙公約を特定のカテゴリに分類できる。しかし、このような構造のデータを応答変数 (ターゲット変数) として機械学習[8]を行う場合、誤分類の影響を無視できなくなる。上神・佐藤 (2009) は、C4.5 アルゴリズム[9]を用いて三次元構造のデータを分類し、「対象」の分類において約 80% の再現率を達成した。とりうる値が 34 個であることを考えると非常に良い性能である[10]。しかし、1 つの

6) 大分類を基準にした場合、16 個である。
7) 厳密にいえば、維持 (t)、転換・改革 (w)、その他 (z) 以外にも過去の業績 (x) という項目もある。
8) ロジスティック回帰分析や線形判別分析などの統計的学習を含む。
9) 上神・佐藤 (2009) では J48 アルゴリズムと表記されているが、基本的には、C4.5 アルゴリズムの Weka 版である。C4.5 は ID.3 の拡張版である。
10) 二項変数をターゲットにした場合、すべて 1 あるいは 0 に予測すると再現率は必ず 50% 以上となるが、取りうる値が 34 個の場合は約 3% であり、80% の再現率は非常に高いといえよう。

要素を誤って分類した場合、とりうる値が 33 個に減少することも同時に意味する。たとえば、「国民 (41)」を「低所得者 (57)」や「外国人 (58)」に分類すると、公約の意味は大きく変わり、元の情報は完全に失われる。再現率が 100% に近い場合、このような問題は生じないが、再現率が低くなるにつれて、得られる予測の信頼性も同様に低下する。

上記では、政治的テキストデータ分析の変遷、並びに政治的テキストデータの構造に関する議論をレビューしてきた。政治的テキスト分析において焦点になるのは、テキストデータの分類の「効率化」と構造の「最適化」である。そこで、第 3 節では、選挙公約の分類手法として多用されてきたヒューマン・コーディングの有効性と限界について、第 4 節では、誤分類の影響を最小限に止めるデータの構造について議論したい。

3　テキストデータのコーディング方法

3.1　ヒューマン・コーディング

本節では、テキストデータの代表的なコーディング方法を 2 つ紹介する。先行研究で用いられてきた方法は、ヒューマン・コーディングである。CMP (Budge et al., 2001; Klingemann et al., 2006; Volkens et al., 2014) のマニフェストデータ[11)]や 品田 (1998, 2006) の選挙公約データのコーディングは、厳格なコーディング・ルールを熟知するコーダーによって行われた。しかし、熟練したコーダーへの業務委託には、ある程度のコストが伴う。たとえば、CMP のコーディング・ルールは数十ページに上るため、コーダー一人の養成には数ヶ月を要する[12)]。分類基準の統一性を担保した上で、コーディングミスを最小化するためには、コーダーの養成に時間的コストを支払うことはやむを得ない。さらに、莫大な量のデータをコーディングするためには、一つのデータセットにつき、数人のコーダーが必要になる。つまり、少数の熟練コーダーによる分類作業に至るまでに、コーダーのトレーニング期間、コーディング期間などの「時間的コスト」と、コーダーを雇用するための「金銭的コスト」を負担しなければならない。これらのコストを負担できるのは大規模なプロジェクトに限定され、一研究者が単独でテキストを分類、分析することは難しい[13)]。

一方、同じヒューマン・コーディングであっても、クラウドソーシングを用いることで、時間的コストと金銭的コストを大幅に削減できる[14)]。世界的なクラウドソーシングである

11) データは https://manifesto-project.wzb.eu/ において入手可能である。
12) CMP データの有用性については，Gemenis (2013) が詳しい。
13) コーダーによる分類の信頼性については、Mikhaylov et al. (2011) が詳しい。
14) Benoit et al. (2016) は、クラウドソーシングを用いたテキスト分析について詳述している。

「AMT」、日本の「Yahoo! クラウドソーシング」を利用することで、数千 ～ 数万人のクラウド利用者に選挙公約のコーディングを委託できる。さらに、クラウドソーシングの場合、熟練コーダーへの作業委託に比べて廉価である。

しかし、“素人”コーダーへの業務委託は、コーディングの正確性を低下させるかもしれない。とくに、一つのテキストに複数の意味が包含される場合、コーダーの間に解釈の余地が生まるため、分類基準が不安定になり、コーディング結果を精査しなければならない。クラウドソーシングにおいても、高い報酬を担保することで、クラウド利用者を訓練することは可能である。ただし、その場合は、上記と同様、コストの問題が生じる。したがって、クラウドソーシングでは、コーダーは “素人”であることを想定しなければならず、分類の正確性の低下は避けられない[15)]。

3.2 機械によるコーディング

もう一つの方法は、機械学習アルゴリズムに基づくテキストの分類である。機械学習によるテキストの分類は、一定数のトレーニング・データを用意した上で、そのデータのパタンをパソコンに学習させ、新たなテキストデータが入力された際、そのテキストデータがどのカテゴリに属するかを予測する。ここで用いるトレーニング・セットは、基本的に人力で構築する[16)]。機械学習のメリットはコストの削減にある。トレーニングセットの構築に時間を費やす必要があるものの、学習はパソコンが行ってくれるため、人力によるコーディングより効率が良い。学習時間は、ハードウェア環境と機械学習アルゴリズムによって数分から数日かかり、ばらつきがある。しかし、いずれにしてもヒューマン・コーディング比べて、非常に短い時間でテキストを分類できることに変わりはなく、依然として機械によるコーディングの有用性は高い。近年、Amazon Web Services や Microsoft Azure などのクラウドプラットフォームを利用して並列処理を行うことも可能であり、これらのサービスを利用することで個人用コンピュータより数十倍の速度でテキストデータのパタンを学習させることができる。また、個人でコーディングを行うにしても、機械学習アルゴリズムを実装するコストは大幅に低下しており、手軽に分類器を生成でき

15) 筆者独自の調査によると “素人”コーダーが熟練したコーダーと同じようにコーディングする割合は、5 割に満たない。とくに、1 つの選挙公約に 2 つ以上の内容を含む場合、コーディングの精度が著しく低下した。

16) クラスター分析、k 近傍法、自己組織化写像など、教師なき学習アルゴリズムでケースをクラスタリングした後、ラベルを付ける方法もあるが、最終的に人力が必要とされることに変わりはない。

るため[17]、金銭的コストの負担は小さい。さらに、分類作業をパソコンに一任するため、ヒューマン・コーディングで起こりがちな入力ミスは生じ得ない。

一方、機械学習にもデメリットや限界はある。機械学習アルゴリズムによって生成される分類器のパフォーマンスは、トレーニング・セットの質に大きく依存する。したがって、少なくともトレーニング・セットを構築する段階では、コーディング・ルールを熟知するコーダーが必要であり、多少なりともコーダーの雇用に伴う時間的コストと金銭的コストを負担しなければならない。

しかし、一度熟練したコーダーによってトレーニング・セットが構築されれば、その後の分析ではコーダーを雇用する必要はなくなる。これは、機械学習によるテキストの分類がヒューマン・コーディングの代替手段であることを意味する。つまり、テキストデータの機械的分類は、熟練コーダーによってコーディングされたデータから規則性を発見し、熟練コーダーによる分類結果に出来る限り近づけることを目的とする[18]。質の高いトレーニング・セットが構築されたら、その後はデータを即時に分類できる。さらに、分類作業を積み重ねることで、トレーニング・セットが増え、分類の精度は高まる。以上を鑑みると、機械によるコーディングは、ヒューマン・コーディングに比べて、依然として有効な手段であるといえよう。

4　選挙公約の文法

本節では、先行研究と言語学の知見に基づき、機械的処理に適した選挙公約データの構造を提示し、その構造を「選挙公約の文法 (Grammar of Manifesto；以下、GoM と略す)」と称する。生成文法 (generative grammar) における変形文法 (transformational grammar) の枠組み (Chomsky, 1957) から選挙公約を見ると、選挙公約に記されている文章を表層構造 (surface structure)、その文章の意味や主張を深層構造 (deep structure) として理解できる。深層構造は、複数の表層構造に変形できる。たとえば、「私、りんご、食べ、た」という深層構造は、「私はりんごを食べた」「りんごは私に食べられた」「りんごは私を食べた」「私はりんごに食べられた」という表層構造に変形できる。このように深層構造の情報が不足していると、表層構造への変換が不完全になりうる。

しかし、選挙公約は目標が明確であるため、深層構造から表層構造への変形は不完全な

[17] 統合されたパッケージとしては、クロスプラットフォームの Weka、R の `caret` パッケージ、Python の `scikit-learn` などがあり、個別のアルゴリズムのパッケージも多数公開されている。

[18] しかし、これは機械学習の分野全般に適用されることではない。たとえば、推薦システムでは、人間では普段感知できないパタンを発見、予測することを目的とする。

ものにならない。各候補者の目標を再選と仮定すれば (Mayhew, 1974)、候補者は選挙公約で有権者の利益になることを述べる。選挙公約の文章は、「有権者の利益に資する政策への言及」という規則性を有しているため、深層構造から表層構造への不完全な変換は生じにくい。選挙公約を生成文法の枠組みから捉えると、深層構造から表層構造へ不完全な変換によって誤解を招くことはなく、したがって深層構造のみを分析の対象にしても問題は生じない。また、深層構造は、単純な構造であるため、機械的分類にも適合する。以下では、深層構造を構成する要素である「公約素」について詳細に解説する。

4.1 公約素の構造

本稿では、構造主義の知見を援用し、選挙公約を交換可能（exchangeable）な意味を有する最小単位の集合として捉え、これを「公約素」と呼ぶ。ソシュールによって提唱された構造主義 (Saussure, 1916) は、文化人類学者であるレヴィ＝ストロースによって体系化された (Lévi-Strauss, 1949, 1962)。構造主義では、ある対象を複数の要素に分解し、分解された要素の組み合わせや、要素間の関係に注目する。構造主義と選挙公約データ分析の間には、一種の共通項を確認できる。具体的には、選挙公約を「対象」「内容」「方向」の要素に分解したことが、それに該当する。

三次元構造の選挙公約データは、誤分類の影響を多分に受けるため、改善の余地がある。誤分類の影響を低減させるためには、可能な限り要素を分解する必要がある。たとえば「対象」を誤って分類し、「内容」と「方向」を正しく分類すると、3 分の 2 の情報が残る。それに対して、要素を数十個に増やすと、誤分類による影響を軽減できる。

選挙公約の内容を詳細な要素に分解する場合、どのレベルまで分解するかが問題になる。本稿では全ての要素を「公約素」に分解する。言語学における音素や形態素は、一つが変わるとその意味が変わる単位を指す。その点は、公約素も同様である。

既存のコーディング方式は、たとえば、データ内の 120 番目候補者の 2 つ目の選挙公約「子ども手当を実現する」を

$$
\begin{aligned}
M_{120,1} = 43\text{v}3\text{w} = \{&\textbf{対象} = \text{生活者 (43)},\\
&\textbf{内容} = \text{少子高齢化対応 (v3)},\\
&\textbf{方向} = \text{改革 (w)}\}
\end{aligned}
$$

の三次元ベクトルで表現する。これをさらに分解すると、以下のように表現できる。

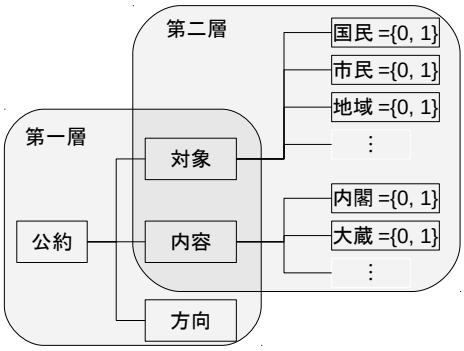

図 1　公約素の層構造

$$
\begin{aligned}
M_{120,1} = \{\text{対象} = \{\ &\text{国民・民意} = 0, \text{市民} = 0, \text{生活者} = 1,\\
&\text{地域公約} = 0, \text{有権者} = 0, \ldots\},\\
\text{内容} = \{\ &\text{保育} = 1, \text{福祉} = 1, \text{交通} = 0, \ldots\},\\
\text{方向} = \{\ &\text{維持} = 0, \text{推進} = 1, \ldots\}\}
\end{aligned}
$$

本稿で提案する方法は、既存の方法と同様に、選挙公約を三次元ベクトルで表現する。しかし、2 つの方法は、ベクトル内の要素も同様にベクトルで構成するか否かで異なる。本稿で新たに提案する方法は、「対象」と「内容」のベクトル内の要素も同様にベクトルで構成する。公約素とは範囲、性別、年齢、所得などの要素を指し、公約素の値が変化すると、その公約の意味も変化する。たとえば、範囲が全体から地域になると、全国を対象とする政策の公約から、特定地域 (選挙区) のみを対象とする政策の公約に意味が変化する。また同様に、所得が全体から低所得になると、低所得者を対象とする政策の公約になる。各公約素が取りうる値は 0 と 1 のいずれかである。ターゲット変数を可能な限り単純化することで、自動分類のパフォーマンスを向上できる。ただし、「各公約素は必ず一つの値のみをとる」との条件を満たす必要がある。つまり、公約素内の要素は相互に排他的でなければならない。

また、公約素は層 (layers) 構造で構成される (図 1 を参照)。「対象」「内容」「方向」から成る次元を「第一層」とすると、第一層内の各要素は「第二層」になる。第二層には、範囲、性別、年齢、所得などの要素が含まれる。この層構造は、無限大に拡張することも、特定要素に対してのみ拡張することも可能であり、柔軟性が高い。図 1 を例にとると、「対象」「内容」「方向」の全てが第二層になるのではなく、「対象」と「内容」のみが第二層になり、「方向」は下位の次元をもたない構造になる。層構造は、文脈に応じて調

節できるものの、次元が増えるとデータのサイズが大きくなるため実用性が劣るため、二層構造が適切であろう[19)]。

4.2 GoM 構造の長所

第 2 節において、既存の選挙公約データは、1 つから 3 つの変数で構成されることを確認した。他方で、第 4.1 小節において提示した GoM 構造のデータは、数十の変数で構成されており、既存のデータ構造に比べてサイズが大きい。したがって、機械学習を用いて GoM 構造のデータを分析する場合、多大な時間的コストを支払わなければならない。上神・佐藤 (2009) は、「対象」「内容」「方向」を応答変数（ターゲット変数）としたため、分析は 3 回で済んだが、GoM 構造データはこの数倍の分析時間を要する。しかし、機械学習の分析が長時間に及んでも、数ヶ月が費やされるヒューマン・コーディングに比べれば短い時間で分類は終わる。さらに、分析を支えるハードウェアの進歩によって分析時間は大幅に短縮される[20)]。

また、GoM 構造のデータは、分析所要時間の長期化という短所を相殺できる複数の長所を有する[21)]。第一に、GoM 構造のデータは、既存のデータ構造に比べて情報量が多く、既存の構造では表現できない値を表現できる。たとえば、選挙公約の対象を「障害をもつ女性」と表記する際、品田 (1998, 2006) の三次元構造データでは、これを的確に表現できる対象カテゴリがないため、「女性」と「障害者」として扱い、それぞれに 0.5 の重みを付けるしかない。一方、GoM 構造データでは $\{$ 女性 $= 1,$ 障害者 $= 1\}$ と表記するだけで済む。このように第二層内の変数を組み合わせることで多様な意味を表現できる。

第二に、GoM 構造のデータは、カテゴリの追加と削除を柔軟に行える。コーディングの最中に新しいカテゴリを追加する際、これまでのコーディング方法では、カテゴリの内容を見直す必要があった。コーディングの過程で必要に応じてカテゴリを追加すること

---

19) 選挙公約の通時的・共時的分析のためには、一定程度の一貫性を保つ必要があり、文脈に応じて、公約素を設定し直さなければならない。たとえば、政策の対象として北朝鮮から脱出した人（脱北者）を第二層に含むのは、韓国では有効であるが、日本では意味をもたない。逆に在日韓国人・朝鮮人を第二層に含むのは、日本では有効であるが、韓国では意味をもたない。

20) データ構造が異なるため、上神・佐藤 (2009) との比較は難しいが、決定木モデルに基づく分類器 (C5.0 アルゴリズム) の場合、1 つの項目の学習に約 10 分の時間を要した。計 54 項目を学習させた本稿のモデルでは、最長 540 分、9 時間を要する。これは、C5.0 より単純なアルゴリズムである C4.8 を用いて 3 項目の学習に 36 時間を要した上神・佐藤 (2009) に比べれば、非常に短い時間である。これは 0 と 1 の二値変数のみを用いる GoM 構造固有の長所である。むろん、ハードウェアや最適化アルゴリズムの発展に起因する分析時間の短縮も無視できない。

21) GoM 構造のデータは、誤分類による影響を最小化できる。この点は、第 4.1 小節で述べたため、ここでは言及しない。

表 1　変換前のターゲット変数

ID	対象	内容	方向
1	国民, 女性	厚労, 内閣	改革

で、カテゴリ数が増え、逆に各カテゴリに属するケースは少なくなるため、機械学習のパフォーマンスが悪くなる[22)]。しかし、GoM 構造のデータは、カテゴリを削除してもコーディングを見直す必要はないため、相対的にカテゴリの追加・削除の制約を受けない。

第三に、GoM 構造のデータは、過去の選挙や異なる種類の選挙における選挙公約データをトレーニングセットとして活用できる。たとえば、適切にコーディングされた 2011 年衆議院議員選挙の選挙公約データをトレーニングセットとして、2014 年参議院議員選挙の選挙公約データを分類できる。すなわち、「対象」「内容」などの要素が 2 つの選挙において一致すれば、別途の処理を施さずに、2011 年衆院選の選挙公約データをそのままトレーニングセットとして用いて、2014 年参院選の選挙公約データを分類できる。ただし、新しい要素が追加されたり、削除されたりすると、そのままトレーニングセットとして用いることは出来ず、改めてコーディングし直さなければならない。

第 5 節では、本節で紹介した GoM 構造のデータを、どのようにして機械的に分類するのか、詳述する。

5　機械学習による選挙公約の分類

5.1　データ

本稿では、品田 (1998, 2006) の基準に従い分類された 2009 年衆議院議員総選挙の選挙公約データ[23)]を用いて、GoM 構造のデータの有効性を確認する。本稿で用いるデータは、品田 (2010) で用いられたデータと基本的に同じである。相違は、分割された公約を一つの文章にまとめたため、一つの文章に含まれる内容の数が多くなるという点にある。しかし、GoM 構造のデータでは、一つの公約内に複数の情報を格納できるため、その点は問題にならない。むしろ、一つの公約の内容が長くなることは、情報量の増加を意味し、より正確な選挙公約の分類に寄与する。

さらに、本稿と品田 (2010) のデータは、ターゲット変数の構造が異なる。品田 (2010)

22) あるいは、学習に用いるトレーニングセットのサイズが大きくなる。
23) データの詳細については品田 (2010) を参照していただきたい。

表 2　変換後のターゲット変数

ID	対象			...	方向			
	国民	市民	女性	...	維持	改革	その他	業績
1	1	0	1	...	0	1	0	0

のデータのターゲット変数は、5 桁の英文字と数字の組み合わせによって表現される。これを「対象」「内容」「方向」に分割すると、表 1 のようになる。そして、品田 (2010) のデータを GoM 構造に変換したものが表 2 である。表 2 からも明らかなように、GoM 構造のデータでは、「対象」「内容」「方向」を全てダミー変数に変換し、端的に表現する。

また、形態素解析は MeCab[24] を用いた。ここでは、抽出された形態素を全て使うのではなく、名詞、動詞、形容詞、数字のみを用いる[25]。形態素解析後のデータは、各行が一つの公約、各列が形態素の出現回数である。最終的に完成したターム文章行列 (term-document matrix) のサイズは 16603 × 7831 である。

5.2　前処理

本格的な機械学習を行う前に，データの前処理 (pre-process) を施すことで、より効率的な分析を行える。生の選挙公約データを形態素解析し、形態素頻度マトリックスを作ると、データのサイズは

$$\text{選挙公約の数} \times \text{全選挙公約内における形態素の数}$$

となり、数百メガバイトから数ギガバイトに達する。このようなサイズのデータを個人用のコンピュータで分析することは難しいため、円滑な分析のために、次元削減 (dimension reduction) の一般的手法である主成分分析[26]により変数の数を減らした。主成分分析を用いた次元削減は、データのサイズを小さくすることによる訓練時間の短縮というメリットがある。加えて、各変数に含まれているノイズを減らす効果も期待できる。理論的に変換されうる主成分の最大数は元の変数の数であるが、それでは次元削減の意味がないため、分散の 70% を基準として情報量の少ない主成分をカットする。この作業によって、変数は 7831 個から 3048 個となり、約 5 分の 2 のサイズまでデータを削減できた。

[24] MeCab の詳細は http://taku910.github.io/mecab/ を参照していただきたい。
[25] 形態素解析には Python 3.5.1 と MeCab ライブラリを使用した。
[26] 他の一般的な次元削減手法としては特異値分解 (Singular Value Decomposition; SVD) があるが、本稿では検討対象外とする。

また本稿では、テキスト分析において一般的である TF-IDF による重み付けは用いない。これは、ある単語あるいは形態素に重みを付けるものであり、頻繁に出現する単語には小さい重みを、稀に出現する単語には大きい重みを付ける。この重み付けはテキスト分析において有効であるが、主成分分析を行う場合、結果的に同じ行列が返される。そのため、本稿では単に主成分分析のみを行う。

5.3 モデル

本稿では、機械学習のモデルを用いて、選挙公約データを予め設けられたカテゴリに分類する。機械学習のモデルは、数十種類、その派生型まで含むと数百種類に上る。これまで、日本の選挙公約を用いて機械学習を行った代表的な研究は、上神・佐藤 (2009) である。この研究は、C4.8 アルゴリズムに基づく決定木 (Decision Tree) モデルを用いて分類を行った。決定木モデルは、比較的計算速度が早く、適切に枝刈り (pruning) を行えば[27]、過学習 (overfitting, overtraining) を回避しながら優れたパフォーマンスを示す。したがって、本稿においても決定木モデルを用いる。ただし、本稿では性能を向上させるために、ブースティングの一種である AdaBoost (エイダーブースト；Adaptive Boosting)[28]を行う C5.0 アルゴリズムを用いる[29]。

選挙公約を文章単位で分析すると、形態素頻度行列 (term frequency matrix) のほとんどのセルは 0 になり、無分散に近い列も多く出てくる。この問題は、主成分分析による次元削減、TF-IDF のような重み付け、Box-Cox 変換などで、ある程度防ぐことができる。また、決定木モデル (Decision Tree；DT) を用いて無分散に近い列を多く含むデータをトレーニングすると、過度に枝刈りをする可能性がある。このように、情報量に比べて変数が多くなると「次元の呪い (curse of dimensionality)」に陥る恐れがある。その時に効果的な分類器のモデルは、ランダム・フォレスト (Random Forest；RF) やニューラル・ネットワーク (Neural Network；NN) である (Lantz, 2013)。これらのモデルは決定木モデルに比べ、分類器の生成まで長い時間を要するが、相対的に優れたパフォーマンスを示す。そのため、本稿ではこの 2 つのモデルによる分類結果も評価する。

---

[27] 上神・佐藤 (2009) で用いられた C4.8 アルゴリズムは C4.5 アルゴリズムと同様であるが、自動的に枝刈りを行う点で異なる。

[28] ブースティングは集団学習の枠組みの一つであり、弱い分類器を複数生成し、予測と評価を繰り返しながら各分類器の重みを調整する。ブースティング以外にもバギング (Bugging) やランダム・フォレスト (Random Forest) などがある。

[29] C5.0 は、C4.5 に比べ、ブースティングが可能な点以外にも、速度やメモリーの側面において長所を有する (Wu et al., 2008)。

表 3　分類器のパラメータ設定

分類器	パラメータ
DT	`C50` 関数の初期設定値
RF	決定木の個数 = 500, 決定木内の特性の数 = $\sqrt{N_{\text{features}}}$,
	決定木の最大深度 = 20
NN	隠れレイヤーの数 = 2,
	各レイヤー内のニューロン数 = {200, 200},
	活性化関数 = 双曲線正接 (tanh)

分類器は、いくつかのパラメータで構成される。表 3 は各分類器のパラメータを示している。ランダム・フォレストならツリーの個数やツリー内の特性 (features) の数、ニューラル・ネットワークならニューロンの個数やレイヤーの数がパラメータになる。これらのパラメータの最適値を求めるアルゴリズムは現在のところ確立されておらず、グリッドサーチ (grid search)[30] で最適なパラメータを求めるのが一般的である。しかし、本稿はグリッドサーチによるパラメータの調整は行わない。なぜなら、決定木モデルなら個人用のコンピュータでかろうじてグリッドサーチでパラメータの推定はできるが、ランダム・フォレストやニューラル・ネットワークでは推定に莫大な時間を要するためである。したがって、第 6 節で示す分類結果の評価は、改善の余地があることを断っておきたい。むろん、ランダム・フォレストにおけるツリー内の特性の数のように「特性の数の平方根[31]」などといった一般的に用いられる基準がある場合は、それに従う。

最後に、分析環境を簡単に紹介する。分析で用いるハードウェアは Apple 社の Macbook Pro の 2016 年モデル[32]、使用ソフトウェアは R 3.3.2[33]、R パッケージは C5.0 の場合は `C50`、ランダム・フォレストとニューラル・ネットワークは `h2o` パッケージを使用する[34]。

第 6 節では、上記の環境の下で分析を行い、分析の結果と分析の評価を行う。

30) 調整したいパラメータの値の全ての組み合わせに対して分類器を作成し、評価する作業を指す。

31) たとえば、形態素頻度行列の列数が 1 万の場合、各ツリー内の特性は一般的に 100 と設定する。これは絶対的な基準ではなく、パラメータの調整や生成された乱数によってパフォーマンスは上下しうる。

32) Intel Core i5 3.1 GHz (2 cores), 16GB RAM

33) 既存の R の高速化バージョンである Microsoft R Open 3.3.2 を使用した

34) `C50` の詳細は、`https://cran.r-project.org/web/packages/C50/index.html`、`h2o` の詳細は、`https://cran.r-project.org/web/packages/h2o/index.html`を参照していただきたい。

表 4　混同行列の例

	Positive(実際)	Negative(実際)
Positive(予測)	True Positive(TP)	False Positive(FP)
Negative(予測)	False Negative(FN)	True Negative(TN)

6　分類結果の評価

6.1　分類結果の評価方法

機械学習による分類の結果を評価する方法や指標は複数ある[35)]。本稿では、データをトレーニングセットとテストセットに分割して交差検証 (cross validation) を行い、分類の結果を評価する。具体的には、全体のデータから無作為に 8 割を抽出 (トレーニングセット) し、このデータを用いて分類器の訓練を行う。つづいて、残りの 2 割のデータ (テストセット) を用いてターゲット変数の予測を行い、実際のデータと比較することで、分類器の性能を評価する[36)]。

モデルの評価は、Cohen の κ 統計量を用いて行う (Cohen, 1960)。κ 統計量の詳細は Cohen (1960) に譲り、ここでは計算方法の簡単な紹介に止める。表 4 は分類結果をまとめる混同行列 (confusion matrix) の例である。この表に基づき、κ 統計量は以下の式 1 のように算出する。

$$
\begin{aligned}
p_0 &= \frac{TP + TN}{N} \\
p_e &= \left(\frac{TP + FP}{N} \times \frac{TP + FN}{N}\right) + \left(\frac{FN + TN}{N} \times \frac{FP + TN}{N}\right) \\
\kappa &= \frac{p_0 - p_e}{1 - p_e}
\end{aligned}
\tag{1}
$$

理論上、κ は-1 から 1 までの値をとり、κ が高いほど優れた分類モデルであると解釈する。ただし、「優れた」モデルと評価できる絶対的な基準は設けられていないため、本稿

35) 本稿で用いる κ 統計量以外にも F-measure や AUC などがある。

36) 広く使われている交差検証の方法としては k-fold 交差検証、Leave-one-out (Loo) 交差検証がある。しかし、k-fold 法の場合、交差検証のために費やされる時間が本稿の単純な交差検証より k 倍かかり、Loo の場合は約 16000 倍の時間を要する。高速コンピューティング環境では可能であるが、一般的な環境では時間的な制約がある。したがって、本稿では単純な交差検証によるモデルの評価に止める。

表 5　Viera and Garrett(2005) による κ 統計量の評価基準

κ の範囲	評価
$\kappa < 0.00$	偶然による一致より低い一致度 (Less than chance agreement)
$0.01 < \kappa < 0.20$	ほとんど一致していない (Slight agreement)
$0.21 < \kappa < 0.40$	やや一致している (Fair agreement)
$0.41 < \kappa < 0.60$	まあ一致している (Moderate agreement)
$0.61 < \kappa < 0.80$	かなり一致している (Substantial agreement)
$0.81 < \kappa$	ほぼ完全に一致している (Almost perfect agreement)

ではViera and Garrett (2005) によって提示された κ 統計量の評価基準を用いる。表 5 は κ 統計量の評価基準を示しており、0.61 以上の値をとる場合に、高いパフォーマンスの分類器と評価する。

また、分類結果を直感的に示す指標として的中率 (accuracy) が用いられる (e.g., 上神・佐藤, 2009)[37]。しかし、本稿では的中率を積極的に解釈しない。なぜなら、的中率は真の値の分布によって解釈が異なるためである。たとえば、「10 代 = 1」をとるケースが全ケースの 1% しか占めない 、言い換えれば「10 代 = 0」が全ケースの 99% を占める場合、全ての値を 0 と予測するだけで、的中率が 99% になる。値のバランスが同一でない複数の要素を比較の対象にしうる的中率は、適切な指標とは言いがたい[38]。むろん、的中率が高くなると κ 統計量も高くなる傾向にある。ただし、κ 統計量はエラー率[39] (p_e) を含む指標であるため、必ずしも「高い的中率 = 高い κ 統計量」とはならない。したがって、複数のターゲット変数を同時に比較するときには、的中率ではなく κ 統計量に依拠することで、評価の妥当性を高めることができる。

[37] 的中率は式 1 の p_0 を意味する。詳細は、Sebastiani (2002) を参照していただきたい。
[38] 言い換えると、同一のターゲット変数を用いた複数のモデルの比較の場合は有効な指標となりうる。
[39] 的中率は式 1 の p_0 と一致する。

表 6　各分類器の分類結果の要約

	DT		RF		NN	
	的中率	κ	的中率	κ	的中率	κ
対象						
平均	0.979	0.493	0.985	0.653	0.987	0.767
標準偏差	0.047	0.193	0.032	0.174	0.032	0.138
内容						
平均	0.947	0.568	0.962	0.607	0.965	0.718
標準偏差	0.033	0.156	0.022	0.174	0.025	0.168
方向						
平均	0.878	0.405	0.924	0.560	0.911	0.506
標準偏差	0.117	0.252	0.069	0.220	0.045	0.164
全体						
平均	0.961	0.516	0.973	0.626	0.974	0.728
標準偏差	0.058	0.184	0.037	0.175	0.040	0.164

6.2　分類結果

本小節では、表 3 のパラメータ設定に基づき、実際にデータを分類する。分類器ごとに結果をまとめたものが表 6 である。紙幅の関係上、本稿では、「対象」「内容」「方向」ごとに算出された的中率と κ 統計量の平均値、標準偏差のみを示す。分類結果の詳細については 付録 B を参考されたい[40)]。

本稿では、的中率に基づくパフォーマンスの評価を積極的に行わないことは既述の通りである。しかし、上神・佐藤 (2009) が的中率を基準にモデルの評価を行なっているため、ここでは簡単に比較する。むろん、ターゲット変数が異なること、形態素解析のエンジンが

40) 付録 B の中で κ 統計量が空白になっている箇所がある。これは分類器による予測値が全て 0 あるいは 1 になる場合である。混同行列で 3 つのセルが 0 の場合、κ 統計量を算出できず、2 つのセルが 0 の場合、κ 統計量は必ず 0 であるため、κ 統計量の意味をもたない。κ 統計量が計算不可、0、あるいは非常に低いのはデータのターゲット変数が極端に偏っているケースである (詳細は 付録 A を参照されたい)。選挙公約を用いて政党・候補者の政策位置を多次元尺度法などで推定する際、このようなケースは事前に除去する場合が多く、実践的な意味では大きな問題はないと考えられる。

異なること[41]、次元削減の有無[42]など、本研究と上神・佐藤 (2009) の間には複数の相違があるため、直接的な比較は困難であることを断っておきたい。

上神・佐藤 (2009) は、10-fold 交差検証で的中率を求めた。「対象」「内容」「方向」の的中率は、それぞれ約 80.2、71.0、84.0% である。一方、本稿の分析では、GoM 構造の選挙公約データを用いて分類器を生成し、決定木モデルで約 97.9、94.7、87.8% の的中率、ランダム・フォレストで約 98.5、96.2、92.4% の的中率、ニューラル・ネットワークで約 98.7、96.5、91.1% の的中率を達成した[43]。的中率のみに焦点を絞ると、アルゴリズムによる違いは大きくないことがわかる。実際、ターゲット変数が単純であり、かつ、トレーニング・セットとテスト・セットの特徴が均質である場合、アルゴリズムによる違いは小さくなる。本稿のデータは無作為にトレーニングとテスト・セットを分割したため、このケースに該当する。

一方で、κ 統計量を中心にモデルの評価を行うと、アルゴリズムによるパフォーマンスの改善を確認できる。Viera and Garrett (2005) の κ 統計量を基準に分類結果を評価すると、優れたパフォーマンスを示していることが分かる。ランダム・フォレストやニューラル・ネットワークによる分類結果は、ほとんどの項目において κ 統計量が 0.6 より大きく、分類器として実用化可能なレベルにある。ただし、「方向」のパフォーマンスについては、κ 統計量が 0.6 以下であり、やや低い。この結果は、維持・転換・業績などのカテゴリに分類された形態素が、解析後の主成分分析の段階で多くの情報を失ったことで生じたと考えられる。形態素をそのまま分析に使う方法もあるが、過学習の可能性を排除できない上に、言語学の詳細な知見も必要になる。

本稿の分類器はパラメータのチューニングを行わなかったが、その点を汲みすれば十分に良い結果であるといえよう。より性能の高いコンピュータを用いる、あるいは GPU を用いて並列計算ができれば、グリッドサーチによるパラメータ調整が可能となり、精度の高い分類結果が期待できる。

---

41) 本稿は MeCab を、上神・佐藤 (2009) は GoSen を使用した。

42) 本稿は主成分分析を行い、分散の 70% でカットした。一方、上神・佐藤 (2009) では次元削減を行わず、TF-IDF による重みづけのみを施した。一ケース内の情報量の側面からみると、上神・佐藤 (2009) の方が優れている。しかし、上神・佐藤 (2009) の分析では、分類器生成に長時間を要し、さらには過学習の恐れもある。

43) 各モデルの的中率は、「対象」「内容」「方向」の順に表記している。

7　結論と課題

本稿は、選挙公約を効率的に分類する方法と、その方法に最適なデータの構造を明らかにした。数ある政治的テキストデータにおいて、候補者単位で発行される選挙公約は、政治家や政党に関する多様な情報を包含するデータとして位置付けられる。そのため、選挙公約は、政党や政治家の政策位置を相対的に比較・分析する際、有用なデータである。しかし、選挙公約が包含する情報量の多寡が分析コストの上昇をもたらし、これまで積極的に活用されてこなかった。つまり、選挙公約分析において、データの「質」を優先することと、「分析コスト」の負担とがトレードオフの関係にあった。

しかし、本稿で提示した機械学習の手法を用いれば、数万行に上る候補者単位の選挙公約を大きなコストをかけることなく分類できる。効率的に機械学習を行うためには、分析手法の精緻化もさることながら、用いるデータの性質も同様に重要になる。本稿は、機械学習の便宜性を高めるために、言語学的知見に基づき「選挙公約の文法 (Grammar of Manifesto；GoM)」を提示した。先行研究の選挙公約データは、「対象」「内容」「方向」の三次元で構成されるため、「誤分類による情報の喪失」という脆弱性を抱える。他方、GoM 構造のデータは、先行研究と同様、選挙公約を三次元で表現することに変わりはないが、1 つの次元が複数の次元で構成され、入れ子型の構造になっている点で異なる。この構造のデータは、1 つの公約が包含する情報の喪失を最小限に止め、機械学習のパフォーマンスを高める。

実際、GoM 構造のデータを用いて機械的分類を行うと、既存のデータの構造に比べて高いパフォーマンスを示した。本稿と先行研究は、選挙公約の分析という点で共通点はあるものの、データの構造そのものが異なる。さらに、選挙公約の分類パフォマーンスの評価基準も異なる。先行研究は、主に的中率に基づきパフォーマンスを評価する。しかし、的中率による分類パフォーマンスの評価には限界があるため、本稿では的中率に代わる指標として κ 統計量を用いた評価も同時に行った。分析の結果、的中率を比較してもなお GoM 構造の分類性能は高く、先行研究のデータの構造よりも優れていることが明らかになった。さらに、Viera and Garrett (2005) の基準による κ 統計量の評価においても、GoM 構造のデータは優れたパフォーマンスを示した。

第 3 節で述べたように、GoM 構造は、既存のデータに比べ、柔軟にカテゴリを追加・削除できる。この点から、前回（$t-1$ 期）の選挙公約データで訓練したモデルを用いて、次回（t 期）の選挙公約データを分類できるとの理解も可能である。選挙ごとに公約のカテゴリが異なる場合、ある時点の選挙公約データをトレーニングさせたモデルを、その他

の時点の選挙公約データの予測に用いることはできない。しかし、GoM 構造のデータは、このような問題を回避できる。すなわち、コーディング済みの 2009 年衆院選の選挙公約があれば、2012 年の衆院選や 2016 年の参院選の選挙公約を機械的に分類できる。それぞれの選挙は、選挙の種類や時期によって文脈や位置付けは異なるが、GoM 構造はそのような文脈とは独立している。文脈を考慮する必要がある場合でも、一部の項目をコーディングするだけで、異なる選挙の公約を機械的に分類できる。本稿では、紙幅の関係上、以上のような GoM 構造の長所を経験的に示せなかったため、この点は今後の課題とさせていただきたい。

「代議制民主主義において、有権者は “正しい” 判断の下に、候補者を選択できているのか」、これまで幾度となく問われてきた問題である。有権者の政治的決定は、政治家が政治的情報を発信し、有権者がそれを受信することで下される。そして、政治家と有権者の関係を繋ぐ媒体の一つとして政治的テキストは位置付けられる (Grimmer and Stewart, 2013)。それゆえに、政治的テキストは、有権者の “正しい”判断を規定する媒体の一つであるともいえよう。つまり、政治的テキストを適切に分析する方法の開発は、学術的意義のみならず社会的にも大きな意義をもちうるのである。選挙公約データは Vote Matching にも応用可能であり、GoM 構造が包含する豊富な情報量は有権者と候補者の的確なマッチングを可能にするだろう。本稿で提示した分析手法が、テキスト分析の発展に寄与することを期待したい。

謝辞

本稿で用いた選挙公約データは品田裕神戸大学大学院法学研究科教授から提供を受けたものである。データの使用をご快諾くださった品田先生に、この場を借りて感謝申し上げたい。また、機械学習について貴重なご意見を提供してくださった金丙烈（キムビョンヨル）氏 (NAVER Corp.) に感謝の意を表する。

参考文献

Althaus, Scott L, Jill A Edy, and Patricia F Phalen (2001) “Using Substitutes for Full-text News Stories in Content Analysis: Which Text Is Best?” *American Journal of Political Science*, Vol. 45, No. 3, pp. 707–723.

Beauchamp, Nicholas (2017) “Predicting and Interpolating State-Level Polls Using Twitter Textual Data,” *American Journal of Political Science*, Vol. 61, No. 2, pp. 490–

503.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov (2009) "Treating words as data with error: Uncertainty in text statements of policy positions," *American Journal of Political Science*, Vol. 53, No. 2, pp. 495–513.

——— , Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov (2016) "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data," *American Political Science Review*, Vol. 110, No. 2, pp. 278–295.

Boomgaarden, Hajo G and Rens Vliegenthart (2007) "Explaining the rise of anti-immigrant parties: The role of news media content," *Electoral studies*, Vol. 26, No. 2, pp. 404–417.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum (2001) *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*: Oxford University Press.

Catalinac, Amy (2016) "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections," *The Journal of Politics*, Vol. 78, No. 1, pp. 1–18.

Chomsky, Noam (1957) *Syntatic Sturctures*: Mouton&Co.

Cohen, Jacob (1960) "A Coefficient of Agreement of Nominal Scales," *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46.

Gemenis, Kostas (2013) "What to Do (and Not to Do) with the Comparative Manifestos Project Data," *Political Studies*, Vol. 61, pp. 3–23.

Grimmer, Justin (2015) "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together," *PS: Political Science & Politics*, Vol. 48, No. 1, pp. 80–83.

——— and Brandon Stewart (2013) "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, Vol. 21, No. 3, pp. 267–297.

Hillard, Dustin, Stephen Purpura, and John Wilkerson (2008) "Computer-assisted topic classification for mixed-methods social science research," *Journal of Information Technology & Politics*, Vol. 4, No. 4, pp. 31–46.

Hopkins, Daniel J and Gary King (2010) "A Method of Automated Nonparametric Content Analysis for Social Science," *American Journal of Political Science*, Vol. 54, No. 1, pp. 229–247.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael D. McDonald (2006) *Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in Central and Eastern Europe, European Union and OECD 1990-2003*: Oxford University Press.

Lantz, Brett (2013) *Machine Learning with R*: Packt Publishing, （長尾, 高弘訳,『R による機械学習』, 翔泳社, 2017 年）.

Lauderdale, Benjamin E. and Alexander Herzog (2016) "Measuring Political Positions from Legislative Speech," *Political Analysis*, Vol. 24, No. 3, pp. 374–394.

Laver, Michael, Kenneth Benoit, and John Garry (2003) "Extracting Policy Positions from Political Texts Using Words as Data," *American Political Science Review*, Vol. 97, No. 2, pp. 311–331.

Lévi-Strauss, Claude (1949) *Les structures élémentaires de la parenté*: Presses Universitaires de France, （福井, 和美訳,『親族の基本構造』, 青弓社, 2001 年）.

——— (1962) *La pensée sauvage*: Plon, （大橋, 保夫訳,『野生の思考』, みすず書房, 1976 年）.

Lowe, Will and Kenneth Benoit (2013) "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark," *Political Analysis*, Vol. 21, No. 3, pp. 298–313, Jun.

Mayhew, David R. (1974) *Congress: The Electoral Connection*: Yale University Press, （岡山, 裕訳,『アメリカ連邦議会: 選挙とのつながりで』, 勁草書房, 2013 年）.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit (2011) "Coder Reliability and Misclassification in the Human Coding of Party Manifestos," *Political Analysis*, Vol. 20, No. 1, pp. 78–91.

Monroe, Burt L. and Philip A. Schrodt (2008) "Introduction to the Special Issue: The Statistical Analysis of Political Text," *Political Analysis*, Vol. 16, No. 4, p. 351.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith (2010) "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.," *ICWSM*, Vol. 11, pp. 122-129.

Proksch, Sven-Oliver and Jonathan B Slapin (2010) "Position Taking in European Parliament Speeches," *British Journal of Political Science*, Vol. 40, No. 3, pp. 587–611.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev (2010) "How to Analyze Political Attention with Minimal As-

sumptions and Costs,” *American Journal of Political Science*, Vol. 54, No. 1, pp. 209–228.

de Saussure, Ferdinand (1916) *Le Cours de linguistique générale*: Lausanne, （町田, 健訳,『新訳 ソシュール 一般言語学講義』，研究社，2016 年）.

Sebastiani, Fabrizio (2002) “Machine Learning in Automated Text Categorization,” *ACM computing surveys (CSUR)*, Vol. 34, No. 1, pp. 1–47.

Slapin, Jonathan B and Sven-Oliver Proksch (2008) “A Scaling Model for Estimating Time-Series Party Positions from Texts,” *American Journal of Political Science*, Vol. 52, No. 3, pp. 705–722.

Viera, Anthony J. and Joanne M. Garrett (2005) “Understanding Interobserver Agreement: The Kappa Statistic,” *Family Medicine*, Vol. 37, No. 5, pp. 360–363.

Volkens, Andrea (1992) *Content Analysis of Party Programmes in Comparative Perspective: Handbook and Coding Instructions*: Wissenschaftszentrum.

——— , Judith Bara, Ian Budge, Michael D. McDonald, and Hans-Dieter Klingemann (2014) *Mapping Policy Preferences from Texts: Statistical Solutions for Manifesto Analysts*: Oxford University Press.

Wilkerson, John D and Andreu Casas (Forthcoming) “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges,” *Annual Review of Political Science*.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg (2008) “Top 10 algorithms in data mining,” *Knowledge and Information Systems*, Vol. 14, No. 1, pp. 1–37.

乾健太郎・浅原正幸 (2006)「自然言語処理の再挑戦—統計的言語処理を超えて」,『知能と情報』，第 18 巻，第 5 号，669–681 頁.

猪口孝 (1983)『現代日本政治経済の構図』，東洋経済新聞社.

上ノ原秀晃 (2014)「2013 年参議院選挙におけるソーシャルメディア：候補者たちは何を「つぶやいた」のか」,『選挙研究』，第 30 巻，第 2 号，116–128 頁.

大村華子 (2012)『日本のマクロ政体』，木鐸社，270 頁.

梶原晶 (2014)「国会議員の政策選好としての地方分権改革」,『選挙研究』，第 30 巻，第 2 号，91–104 頁.

小林良彰 (1997)『現代日本の政治過程 : 日本型民主主義の計量分析』，東京大学出版会.

——・堤英敬 (2000a)「選挙公約に関する計量分析 (1)」,『選挙』，第 53 巻，第 1 号，

9–12 頁.

——・—— (2000b) 「選挙公約に関する計量分析 (2)」,『選挙』, 第 53 巻, 第 2 号, 11–15 頁.

——・—— (2000c) 「選挙公約に関する計量分析 (3・完)」,『選挙』, 第 53 巻, 第 3 号, 19–23 頁.

品田裕 (2010) 「2009 年総選挙における選挙公約」,『選挙研究』, 第 26 巻, 第 2 号, 29–43 頁.

堤英敬 (1998) 「1996 年衆議院選挙における候補者の公約と投票行動」,『選挙研究』, 第 13 巻, 89–99 頁.

上神貴佳・佐藤哲也 (2009) 「政党や政治家の政策的な立場を推定するコンピュータによる自動コーディングの試み」,『選挙研究』, 第 25 巻, 第 1 号, 61–73 頁.

小林良彰 (2008) 『制度改革以降の日本型民主主義 : 選挙行動における連続と変化』, 木鐸社.

——・岡田陽介・鷲田任邦・金兌希 (2014) 『代議制民主主義の比較研究』, 慶應義塾大学出版会.

品田裕 (1998) 「¡資料¿選挙公約政策データについて」,『神戸法學雑誌』, 第 48 巻, 第 2 号, 541–572 頁.

—— (2002) 「政党配置—候補者公約による析出」, 樋渡展洋・三浦まり（編）『流動期の日本政治—「失われた十年」の政治学的検証』, 東京大学出版会, 第 3 章.

—— (2006) 「選挙公約政策データについて」,『日本政治研究』, 第 3 巻, 第 2 号, 221–249 頁.

付録 A　ターゲット変数の分布

表 7　ターゲット変数の分布 (1)

	全体		トレーニング		テスト	
項目	0	1	0	1	0	1
対象						
40: その他 (対象なし)	6,377	10,226	5,087	8,195	1,290	2,031
41: 国民、民意	15,683	920	12,541	741	3,142	179
42: 市民	16,586	17	13,269	13	3317	4
43: 生活者	16,579	24	13,263	19	3316	5
44: 地域公約	14,280	2,323	11,430	1,852	2850	471
45: 有権者	16,603	0	13,282	0	3,321	0
46: 庶民	16,587	16	13,270	12	3,317	4
47: 消費者	16,587	16	13,270	12	3,317	4
48: 住民	16,581	22	13,226	16	3,315	6
50: 被災者	16,594	9	13,274	8	3,320	1
51: 高齢者	15,770	833	12,614	668	3156	165
52: 女性	16,466	137	13,175	107	3,291	30
53: 子ども・青少年	14,384	2,219	11,527	1,755	2,857	464
54: 青少年 (有権者)	16,441	162	13,149	133	3,292	29
55: 社会人	16,600	3	13,280	2	3,320	1
56: 障害者	16,439	164	13,152	130	3,287	34
57: 低所得者	16,528	75	13,220	62	3,308	13
58: 外国人	16,597	6	13,278	4	3,319	2
59: 被爆者	16,598	5	13,278	4	3,320	1
61: 労働者	16,584	19	13,266	16	3,318	3
62: 勤労者	16,202	401	12,955	327	3,247	74
63: パート	16,304	299	13,053	229	3,251	70
64: 働く女性	16,567	36	13,258	24	3,309	12

表 8　ターゲット変数の分布 (2)

	全体		トレーニング		テスト	
項目	0	1	0	1	0	1
65: 福祉従事者	16,484	119	13,182	100	3,302	19
66: 中小企業	16,082	521	12,864	418	3,218	103
67: 農漁業	15,793	810	12,628	654	3,165	156
68: 大企業	16,396	207	13,123	159	3,273	48
69: 同和地区	16,602	1	13,282	0	3,320	1
70: 商店街	16,560	43	13,249	33	3,311	10
71: 戦争被害者	16,601	2	13,280	2	3,321	0
72: 社会的弱者	16,473	130	13,172	110	3,301	20
73: ベンチャー企業	16,593	10	13,275	7	3,318	3
80: 産炭地	16,603	0	13,282	0	3,321	0
99: その他 (対象あり)	16,413	190	13,137	145	3,276	45
内容						
a: 内閣	14,426	2,177	11,501	1,781	2,925	396
b: 自治	15,235	1,368	12,184	1,098	3,051	270
c: 安保・外交	15,592	1,011	12,495	787	3,097	224
f: 大蔵	15,339	1,264	12,280	1,002	3,059	262
g: 文科	15,234	1,369	12,192	1,090	3,042	279
h: 厚生	12,885	3,718	10,337	2,945	2,548	773
i: 労働	15,131	1,472	12,107	1,175	3,024	297
j: 農水	15,406	1,197	12,310	972	3,096	225
l: 通産	15,499	1,104	12,405	877	3,094	227
m: 運輸	16,242	361	12,996	286	3,246	75
n: 郵政	16,412	191	13,132	150	3,280	41
o: 建設	15,657	946	12,520	762	3,317	184
q: 環境	16,066	537	12,849	433	3,217	104
r: 政治	14,201	2,402	11,380	1,902	2,821	500
v: その他	14,321	2,282	11,420	1,862	2,901	420
k: 構造改革	16,547	56	13,236	46	3,311	10

表 9　ターゲット変数の分布 (3)

	全体		トレーニング		テスト	
項目	0	1	0	1	0	1
方向						
t: 維持	8,035	8,568	6,424	6,858	1,611	1,710
w: 転換	7,094	9,509	5,644	7,638	1,450	1,871
z: その他	16,531	72	13,223	59	3,308	13
x: 業績	3,218	513	12,872	410	3,218	103

付録 B　分類結果の詳細

表 10　各分類機の分類結果 (1)

	DT		RF		NN	
項目	的中率	κ	的中率	κ	的中率	κ
対象						
40: その他 (対象なし)	0.767	0.504	0.844	0.659	0.841	0.663
41: 国民、民意	0.955	0.497	0.966	0.529	0.981	0.811
42: 市民	0.999	-	0.999	0.400	0.999	-
43: 生活者	0.998	-	0.988	-	0.998	-
44: 地域公約	0.869	0.420	0.903	0.492	0.908	0.634
45: 有権者	-	-	-	-	-	-
46: 庶民	0.999	-	0.999	-	0.999	-
47: 消費者	0.999	-	0.999	-	0.999	-
48: 住民	0.996	0.314	0.998	-	0.998	-
50: 被災者	1.000	-	1.000	-	1.000	-
51: 高齢者	0.971	0.682	0.980	0.745	0.988	0.875
52: 女性	0.991	0.432	0.993	0.448	0.995	0.731
53: 子ども・青少年	0.932	0.713	0.953	0.779	0.965	0.857
54: 青少年 (有権者)	0.988	0.044	0.991	-	0.997	0.783
55: 社会人	1.000	-	1.000	-	1.000	-
56: 障害者	0.992	0.515	0.994	0.610	0.997	0.843
57: 低所得者	0.995	0.331	0.997	0.374	0.995	0.346
58: 外国人	0.999	-	0.999	-	0.999	-
59: 被爆者	1.000	-	1.000	1.000	1.000	-
61: 労働者	0.998	0.285	0.999	-	0.999	-
62: 勤労者	0.981	0.531	0.986	0.568	0.987	0.688
63: パート	0.993	0.835	0.995	0.861	0.998	0.940
64: 働く女性	0.996	0.249	0.996	-	0.996	-

表 11　各分類機の分類結果 (2)

	DT		RF		NN	
項目	的中率	κ	的中率	κ	的中率	κ
65: 福祉従事者	0.995	0.468	0.996	0.598	0.997	0.755
66: 中小企業	0.985	0.738	0.990	0.809	0.992	0.865
67: 農漁業	0.966	0.583	0.977	0.695	0.983	0.804
68: 大企業	0.992	0.712	0.996	0.855	0.998	0.912
69: 同和地区	-	-	-	-	-	-
70: 商店街	0.997	0.284	0.997	-	0.997	-
71: 戦争被害者	1.000	-	1.000	-	1.000	-
72: 社会的弱者	0.993	0.497	0.997	0.665	0.997	0.735
73: ベンチャー企業	0.999	0.500	0.999	0.500	0.999	-
80: 産炭地	-	-	-	-	-	-
99: その他 (対象あり)	0.992	0.702	0.996	0.818	0.995	0.802
内容						
a: 内閣	0.913	0.565	0.941	0.652	0.938	0.723
b: 自治	0.928	0.477	0.948	0.514	0.981	0.811
c: 安保・外交	0.969	0.737	0.973	0.739	0.970	0.771
f: 大蔵	0.948	0.629	0.964	0.695	0.965	0.770
g: 文科	0.950	0.647	0.961	0.674	0.973	0.823
h: 厚生	0.897	0.701	0.936	0.805	0.931	0.813
i: 労働	0.939	0.612	0.960	0.689	0.972	0.834
j: 農水	0.962	0.684	0.972	0.740	0.974	0.795
l: 通産	0.940	0.473	0.951	0.440	0.958	0.659
m: 運輸	0.977	0.415	0.983	0.411	0.989	0.711
n: 郵政	0.995	0.754	0.995	0.774	0.997	0.870
o: 建設	0.960	0.614	0.973	0.689	0.971	0.720
q: 環境	0.975	0.529	0.979	0.474	0.980	0.646
r: 政治	0.918	0.676	0.947	0.775	0.943	0.787
v: その他	0.880	0.433	0.912	0.458	0.905	0.601
k: 構造改革	0.996	0.141	0.997	0.181	0.997	0.153

表 12　各分類機の分類結果 (3)

	DT		RF		NN	
項目	的中率	κ	的中率	κ	的中率	κ
方向						
t: 維持	0.773	0.546	0.865	0.730	0.848	0.697
w: 転換	0.782	0.556	0.864	0.723	0.843	0.678
z: その他	0.996	-	0.996	-	0.996	-
x: 業績	0.960	0.114	0.969	-	0.957	0.306