# オンライン自由記述調査のデータ駆動的分析

## －多様な世論を可視化する分析法の検討－

Data-driven Analysis of Open-ended Online Survey Data: A Study of an Analytical Method for Visualizing the Diversity of Public Opinion

森本　智志
Satoshi Morimoto



〈要旨〉
近年、オンライン調査を利用した世論調査の可能性が模索されている。オンライン調査は大量のデータを短期間に収集できるメリットがあり、その積極的な利用によってマイノリティを含めた多様な世論の可視化も可能になると期待される。しかし、データを収集・分析して可視化する方法論はまだ確立されていない。本研究では、自由記述データをボトムアップに分類するデータ駆動的なアプローチを提案する。自由記述の回答文を事前学習した意味空間に埋め込んだのち、意味空間内での距離に基づいて回答文を分類することで、調査者の主観的判断に依らない回答事項のグループ化を行う。オンライン調査によって取得した模擬データに対して提案手法を適用したところ、世論の分析への応用の可能性が示された。

Recently, the potential of online surveys in public opinion research has been explored. Online surveys have the advantage of collecting large amounts of data in a short period of time, and their use is expected to make it possible to visualize diverse public opinion, including minorities. However, how to collect, analyze, and summarize the data has not yet been established. In this study, we propose the application of a data-driven approach to the analysis of open-ended survey data. We achieve bottom-up grouping of open-ended texts without relying on the subjective decisions of the researcher by embedding texts in a pre-trained semantic space and classifying them based on their distances. We applied the proposed method to data from an online survey. The results demonstrate the potential of our method for public opinion analysis.

## 1．背景

高度情報化が進んで情報端末の個人所有が一般化した現在、世論調査や情勢調査にオンライン調査を利用することで、従来の固定電話や電子メールといった通信手段では現実的に調査が困難になっている若年層を含めた広いカバレッジを実現する方法が議論されている（平田・大隈 2021、大栗 2022)。特に世論は、情報源の多様化とオンデマンド化により今後ますます細分化されると想定され、少数派を含む多様な関心の可視化も世論調査のひとつの重要な役割になるだろう。このとき、調査主体が設問設定してその割合を可視化する従来の方法だけでは、求められるマイノリティの顕在化という課題に対して十分に対処できない。

最も簡便に回答者が関心を持つ事柄を収集する方法は、広いテーマ設定に基づく自由記述である。しかし、自由記述は同じ内容であっても表現に個人差が生じるため、大量のデータの内容を集約することに極めて大きな人的コストが生じてしまう。即ち、様々な揺らぎのある自然言語表現から、類似する意味を持つ内容を適切にグループ化する手法を導入することが、大量の自由記述データを用いた政治的関心の可視化に必須であると言える。

文章の意味分類は、古くから自然言語処理と呼ばれる情報科学の分野において研究が進められてきた（Allahyari et al. 2017)。近年は大量のデータで様々な構造を持つニューラルネットを学習させる方法論が発展し、自然言語で書かれた文章からの意味抽出についてもルールベースの方法だけでなくデータから規則性を獲得する方法が数多く提案されている。大量のデータから多様な意見を抽出するという目的を鑑みると、専門家によるトップダウンの意味ラベリングではなく、データ自体に基づくボトムアップなアプローチの導入が重要だと考えられる。

そこで本研究では、エンベディング（embedding）と呼ばれる機械学習によりボトムアップに構築された意味空間への記述データの配置と、データをその分布に基づいてボトムアップに分類するクラスタリングを組み合わせた分析法を提案する。オンライン調査で取得した模擬データに提案手法を適用し、その有効性を検証する。

## 2．手順

### 2-1. 提案手法の概要

提案手法のフローを図１に示す。大きく二つの技術要素から構成される。一つは、自由記述された文章を数値化する工程である。本研究では文章の持つ意味を、構成する語の分散表現（Bengio et al. 2000）の集合として解釈する。分散表現とは、機械学習により語を固定長のベクトルで表したものである。例えば、300 次元のベクトルであれば[0.79, 0.09, …, 0.53]のような 300 個の実数値で表される。意味の近い語同士の距離が近くなるようなベクトルが学習されるため、分散表現への変換は意味空間に配置することと等しい。分散表現の学習には膨大な文章データが必要となるため、本研究では直接の学習は行わず、公開データにより事前学習した学習済みモデルの意味空間に記述文から抽出した語を配置することにした。

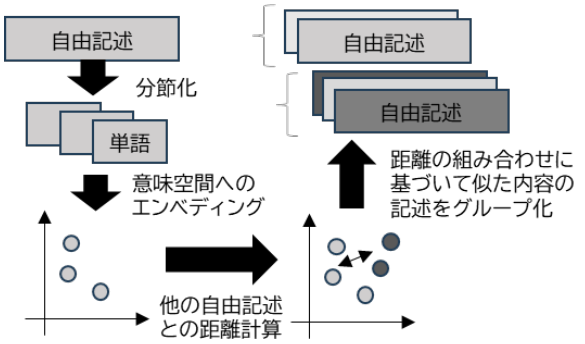


図１．提案手法のフロー

二つ目の工程は回答のグルーピングである。回答同士の距離を、回答に含まれる語の間の意味空間における距離から定義し、距離の組み合わせに基づいて類似した回答をグループ化した。

各工程の詳細は、2-3 及び 2-4 にて述べる。

### 2-2. 模擬調査データ

本研究では、将来的な世論調査における自由記述データの分析を念頭に手法提案を行う。そのため、評価用の模擬データをオンライン調査により取得した。調査は楽天グループ株式会社が提供する「1 万人モーメント」を利用して実施した（https://minijob.rakuten.co.jp/biz/quickorder/)。このサービスでは「楽天超ミニバイト」に登録しているモニターを対象に、簡易なアンケート調査を実施できる。手法評価が目的であるため、一般的な内容の質問で構成した。用いた質問文を表１に示す。

質問文は質問 1 で世界に関する願望、質問 2 ではその理由について自由記述し、質問 3 で実現性に関して評価する段階的な構成とした。調査は

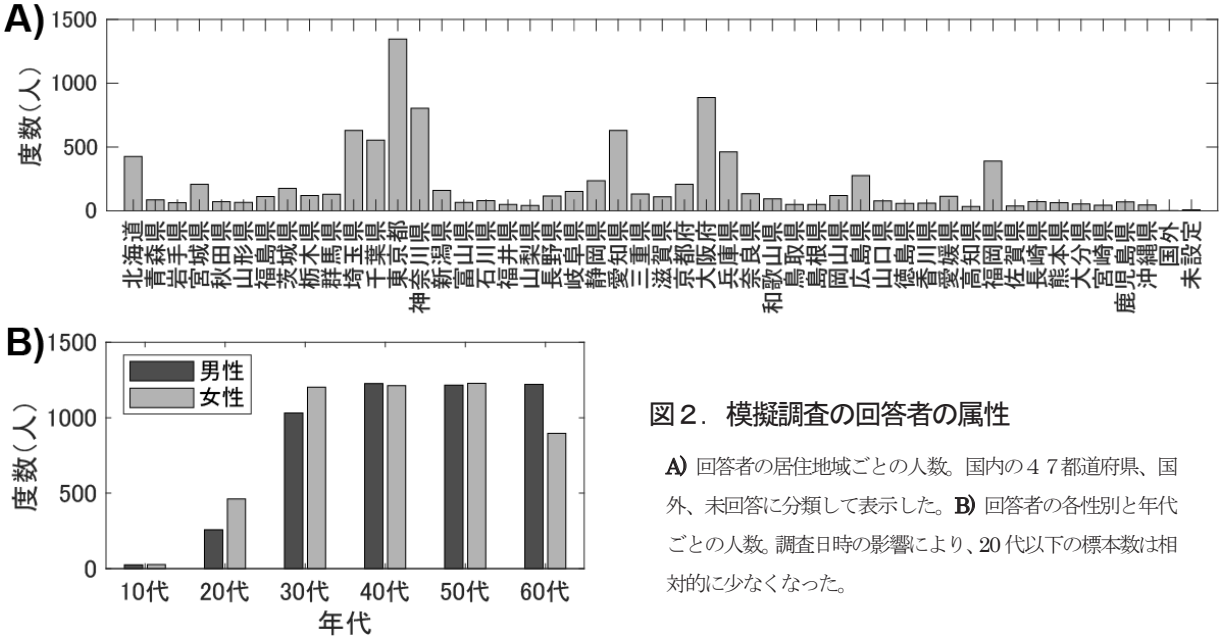


**図２．模擬調査の回答者の属性**

**A)** 回答者の居住地域ごとの人数。国内の４７都道府県、国外、未回答に分類して表示した。**B)** 回答者の各性別と年代ごとの人数。調査日時の影響により、20 代以下の標本数は相対的に少なくなった。

2024 年 1 月 17 日の 14 時 39 分から 21 時 9 分にかけて「あなたの理想の世界」というタイトルでサービスを提供するアプリを介してモニター評価者に提示され、計 10004 名のデータが 10 代から 60 代まで各世代及び男女比が可能な限り均等な標本数になるように自動的に収集された。回答者のデモグラフィックデータを図２に示す。本研究では手法検証を目的とするため、年齢層や性別、居住地域に関して特段の補正は行わずに使用した。

非対面のオンライン調査では、必ずしも有効な回答が得られるとは限らない（三浦 2020）。手法検証の目的であるため虚偽の回答は問題とならないが、無意味文字の入力などの不誠実な回答は、検証に悪影響を及ぼす。そのため、全回答について手作業で確認を行い、無意味な入力を除外した。また、自由記述であるため、回答を思いつかない状況において無回答ではなく「ない」や「なし」という入力を行うケースが散見された。こうした入力も除外対象とした。さらに、質問１に回答せずに質問３に回答しているケースについても、不適切回答と見なして除外した。

**表１．オンライン調査の質問文**

| |
|---|
| 質問１（自由記述） |
| いま、あなたはこの世界がどんな世界になって欲しいと願っていますか？自由に書いてください。（文章でも、単語でも、複数でも可） |
| 質問２（自由記述） |
| なぜそう願っているのか、簡単に教えてください。（文章でも、単語でも、複数でも可） |
| 質問３（選択肢・順不同） |
| その世界は将来実現できると思いますか？<br>□ はい　□ いいえ　□ わからない |

### 2-3. 自由記述文の意味空間ベクトルの算出

自由記述された文章を意味の近さに基づいて分類するため、文章を構成する語を別途事前学習した日本語の意味空間に配置する。分析はすべて Mathwork 社の MATLAB R2023b 上で行い、自然言語処理の分析は Text Analytics Toolbox の関数を用いて実施した。

単語の意味を表現する意味空間へのエンベディングには、Word2Vec（Mikolov et al. 2013b）を用いた。Word2Vec は単語の意味がその周辺の単語から形成されるとする分布仮説に基づき、浅いニューラルネットを用いて分散表現を学習するモデルである。有志によって作成された日本語の学習済みモデルも公開されており、本研究では 2019 年 5 月 20 日時点での日本語版 Wikipedia の本文全文から学習した 300 次元の「日本語 Wikipedia エンティティベクトル」（鈴木ら 2016）を利用した。

今回の分析では、複合語や現在の流行語に対応するため、それらを含有する調査会社から提供された記述回答内のワードリストを最小単位の語として用いた。質問 1 及び 2 への回答のワードリストから検出されたユニークな語（3927 個）それぞ

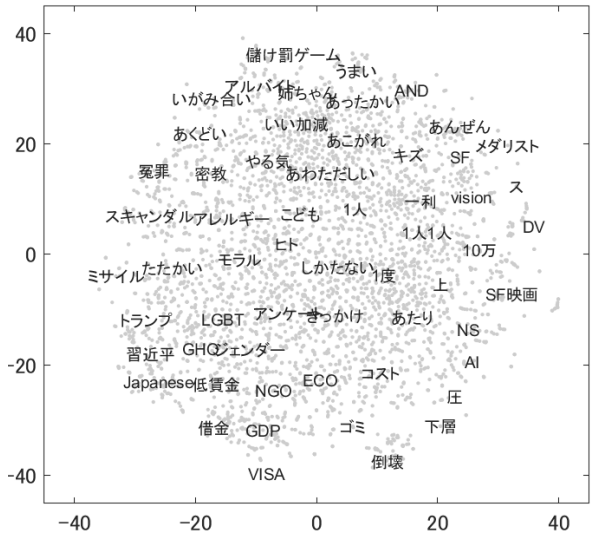


**図３．分析対象の語の意味空間マップ**

全回答文に含まれる語の意味空間ベクトルを、tSNE により二次元でマッピングした。各点は語を示し、距離の近さは意味の近さを近似的に表す。例としていくつかの語について可視化した。

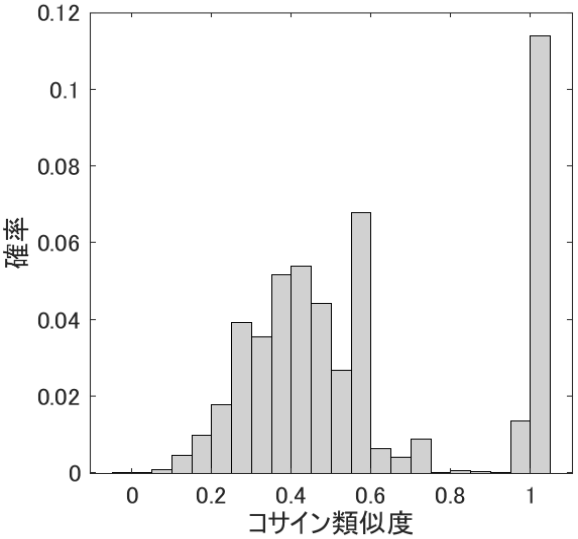


**図４．類似度行列のヒストグラム**

質問 1 に対する全回答文の組み合わせのコサイン類似度の分布を示す。一番右端は類似度が 1、すなわち同じ語の組み合わせが文間の類似度として選ばれたケースを示す。

れについて、Word2Vec の学習済みモデルの意味空間における 300 次元のベクトルを取得した。このとき、normalizeWords 関数を用いて語根に短縮する処理を行ってからベクトルを取得した。分散表現においては、ベクトルの足し算によって意味の足し算ができる加法構成性 (Mikolov et al. 2013a) が成り立つ。そこで複合語や助詞で連結された語に対しては、tokenizedDocument 関数内に組み込まれた MeCab (Kudo 2005) による形態素解析を行ったのち、助詞と文末の「世界」を除外した残りの複数の単語のベクトルの和を算出した。求められたワードリスト内の語の 300 次元の意味空間ベクトルについて、t 分布型確率的近傍埋め込み (t-SNE) により 2 次元で可視化した結果を図３に示す。

### 2-4. 文章間の距離に基づくクラスタリング

似た意味の自由記述文をグループ化するため、意味空間上の距離に基づくクラスタリングを行った。文には複数の語が含まれるため、文間の距離はそれぞれの文内に含まれる語の全ての組み合わせの距離から定義する必要がある。本研究では語間の距離の中から、その分布に基づいて代表値を選んで文間の距離とした。例えば、3 語で構成された文と 5 語で構成された文では 15 の距離が求められるが、文間の距離は 15 の距離の中からその分布に基づいて 1 つ選択した。意味の似た語間の距離ほど短いため、最短距離を示す語間の距離を文間の距離としてクラスタリングすると、文章に共通する語が含まれるほど同じクラスタに分類されやすくなる。即ち、共通表現が反映されやすい定義となり、回答文に頻出しやすい文意を反映しない単語でグループ化されやすくなる。逆に最長距離に基づいてクラスタリングすると、文間に共通しない固有の癖など個性的表現によって同じ文意が細分化しやすくなる。そこで本研究では、語間の距離の中央値を文間の距離の代表値として用いることとした。

意味空間における語間の類似度は、コサイン類似度で定量化した。まず全回答者の自由記述文の組み合わせについて、類似度を算出した。得られた類似度行列のヒストグラムを図４に示す。コサイン類似度の値域は$[-1,1]$であり、同じ語であるとき 1 となる。

今回用いた学習済み意味空間内では、類似度が負の値になる組み合わせはごくわずかで、最小値は-0.057 であった (図４)。そこで文章間の距離$d$は類似度$s$に対し以下のように定義した。

$$d = \begin{cases} 1 - s & (s > 0) \\ 1 & (s \leq 0) \end{cases}$$

続いて、得られた距離行列に対して Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al. 1996) を適用し、

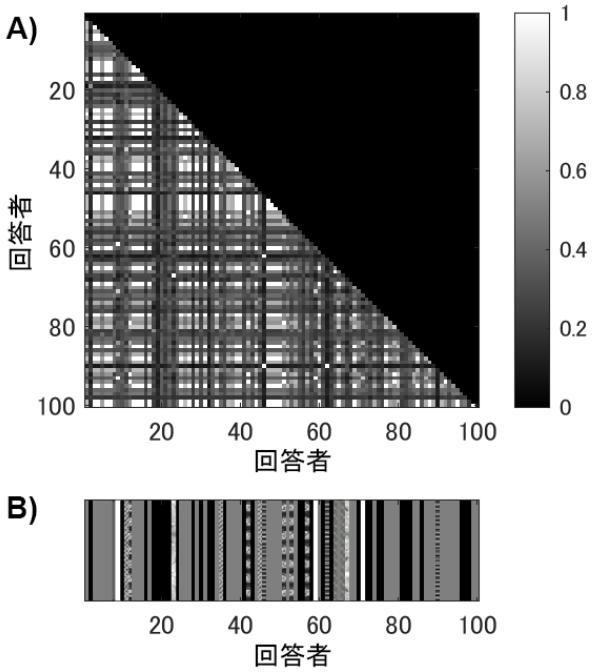


図５．類似度行列とクラスタリングの例

A) 類似度行列の例。100 名分を抜粋して表示した。縦軸と横軸は回答者のインデックスを示す。グレースケールで類似度を表す。回答者の組み合わせの重複を除くため、類似度行列の下三角部分のみ表示した。 B) A の 100 名の回答の類似度行列を用いて DBSCAN により分節化した例。テクスチャの違いでクラスタを表す。黒色部分は最小近傍数以下でどこにも属さなかった回答者を示す。A と比較すると、類似度の高い回答者（回答文）が同じクラスタとしてグループ化されているのがわかる。なお、図の可読性を優先し、DBSCAN のパラメータは実際の分析とは異なるものを使用した（$\varepsilon = 0.4$、最小近傍点数 2）。

グループ化した。距離行列とクラスタリング結果の関係の様子の例を図５に示す。DBSCAN は指定した距離内の密度に基づくクラスタリング手法であり、事前にクラスタ数の指定を必要としない。表現の揺らぎにより意味空間において近い内容の文章は近接していると考えられることから、当手法が妥当であると判断した。DBSCAN のパラメータは類似度行列（距離行列）の分布（図４）を参考に探索半径$\varepsilon$を 0.4、密度の閾値である最小近傍点数を 10 とした。

### 2-4. クラスタの分析

推定されたクラスタについて、その性質を調べるための分析を実施した。

質問 1 と 2 の関係はクラスタの組み合わせの標本数及び条件付き確率で評価した。ただし条件付き確率の計算では、回答のパターンが細分化されるため、クラスタに含まれる標本数の閾値を 10 人と設定した。

クラスタと質問 3 との関係は、質問 3 の回答の「はい」と「いいえ」のみに着目し、各クラスタ内におけるその割合を算出して評価した。

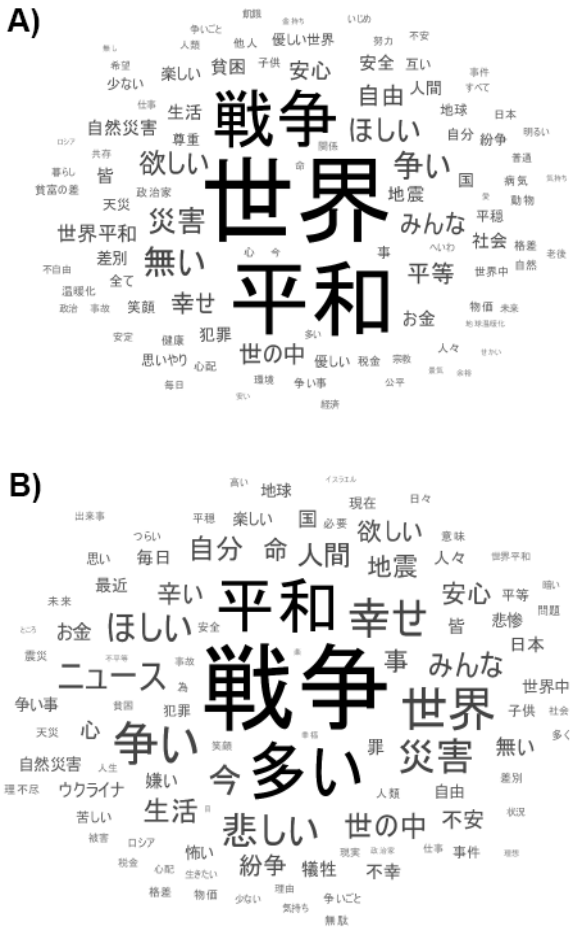


図６．模擬調査データの回答

A) 質問 1 の回答に含まれた語のワードクラウド。B) 質問 2 の回答に含まれた語のワードクラウド。C) 質問 3 の回答分布。無回答・無効回答は除外した。

## ３．結果

### 3-1. 模擬調査

質問 1 および 2 の回答に含まれる語のワードクラウドによる可視化結果を、図６の A と B にそれぞれ示す。質問 1 の 41.4 %において「平和」、21.7 %に「戦争」という単語が含まれており、調査時点で続いているガザ地区における戦争やロシアによるウクライナ侵攻の影響が色濃く出たと考えられる。また調査前に発生した令和 6 年能登半島地震や自民党の派閥資金問題など、時世を反映する語が多く検出された。また質問文に対応して、前処理で除外できなかった文末以外の「世界」という単語も多く検出された。無回答や無効回答の割合は質問 1 が 6.0 %、質問 2 が 6.7 %であった。

質問 3 の有効回答の内訳を図６Ｃに示す。質問 1 で回答した世界の実現性について「はい」の回答が 29.0 %、「いいえ」の回答が 61.0 %、「わからない」の回答が 12.9 %、無回答・無効回答は 7.1 %であった。

### 3-2. 自由記述文の分類

提案手法により、質問 1 と 2 の自由記述の回答文からそれぞれ 32 個、51 個のクラスタが推定された。有効な回答のうち、クラスタに属した回答数の割合は質問 1 が 63.5 %（5829 人分）、質問 2 が 48.8 %（4282 人分）であった。

クラスタ内に含まれる語について、ワードクラウドにより可視化した結果を図７に示す。図６のワードクラウドと異なり、関連する語をグループ化するようにクラスタが形成された。ほとんどの回答が「戦争」または「平和」の語を含む回答であったため、質問 1（図７Ａの#1）も質問 2（図７Ｂの#1、#2、#4、#5）も、それらを含むクラスタが多くを占める結果となった。質問 2 のクラスタには同じ「戦争」の語が主に含まれるクラスタが 2 つ推定されたが（図７Ｂの#1、#2）、これは文章内の他の語との関係で分離したものである。なお、「当」（図７Ｂの#6）や「番」（図７Ｂの#7）といった、一見解釈の困難な語のクラスタも推定された。これらは語の抽出に使用したワードリストによる影響であり（前者は「当然」「当たり前」、後者は「一番」「1 番」に由来する）、提案手法自体には起因しない。

### 3-3. 分類結果に基づく分析

推定された質問 1 と 2 の回答のクラスタについて、両者の組み合わせの標本数を求めた（図８Ａ）。その結果、質問 1 でクラスタ #1 に割り当てられた回答のみが 10 以上の標本数の閾値を満たしたため、当該回答に対してのみ質問 2 のクラスタとの条件付き確率を求めた。質問 2 が質問 1 の回答の理由であることに基づいて、クラスタ間の関係を描写した可視化の例を図８Ｂに示す。「平和」な世界や「戦争」のない世界を希求する理由として、現在「戦争」が発生していることや「平和」ではないことを記述した回答が多数を占めていたため、同義反復のようなクラスタ間関係が目立った。

質問 3 で得られた実現性の評価について、質問 1 のクラスタごとに割合を求めた結果を図９に示す。どのクラスタも実現性については半数以上が悲観的な見通しを回答した。「戦争」「平和」に関連する回答（クラスタ #1、#4）や「幸せ」に関連する回答（クラスタ #2）では実現性がチャンスレベルに近い評価であった一方、「尊重」「平等」（クラスタ #7）や「不安」「ストレス」（クラスタ #6）、「弱者」（クラスタ #5）に関連する回答では悲観視している評価結果が得られた。

## ４．考察

本研究では、将来的な自由記述による世論の可視化への応用を念頭に、文章の意味空間へのエンベディングと文章間の距離に基づくクラスタリングによる回答分類手法を提案した。模擬調査データに対し提案手法を適用し、類似する内容の回答文をボトムアップにグループ化できることを確認した。さらに、推定されたクラスタと他の選択肢回答の関係性を検討する二次的な分析の方向性を示した。これらを用いることで、例えばある政治的要望のクラスタにおける内閣支持の分析など、関心事項の抽出と政策評価の具体的な対応分析が可能となる。自由記述に基づくため、従来の個別設計された調査の回答分析と比べてより関心の高い層に着目した分析となり、特にマイノリティな世論の可視化という側面において有効な手法であると言える。

提案手法の重要な特徴は、入力された記述文章に基づいて半自動的にグループ分けを行うことにある。本研究では学習済の Word2Vec モデルを利用して意味空間へのエンベディングを行ったが、Word2Vec の意味空間は分布仮説に基づいて単語の近接関係のみからボトムアップに学習されたものであり、調査主体による意図の介入の可能性が小さい。また、グルーピング結果に影響を与えうる設定は DBSCAN の 2 つのパラメータのみであり、客観


図７．推定された回答のクラスタ

A) 質問１の回答から推定されたクラスタに含まれた語のワードクラウド。32 のクラスタのうち、全回答者数の 0.1％にあたる 10 人以上の回答を含む 7 のクラスタを人数順で示す。図の上のタイトルはクラスタ番号と標本数を示す。B) 質問 2 の回答から推定されたクラスタに含まれた語のワードクラウド。51 のクラスタのうち、10 人以上の回答を含む 13 のクラスタを人数順で示す。

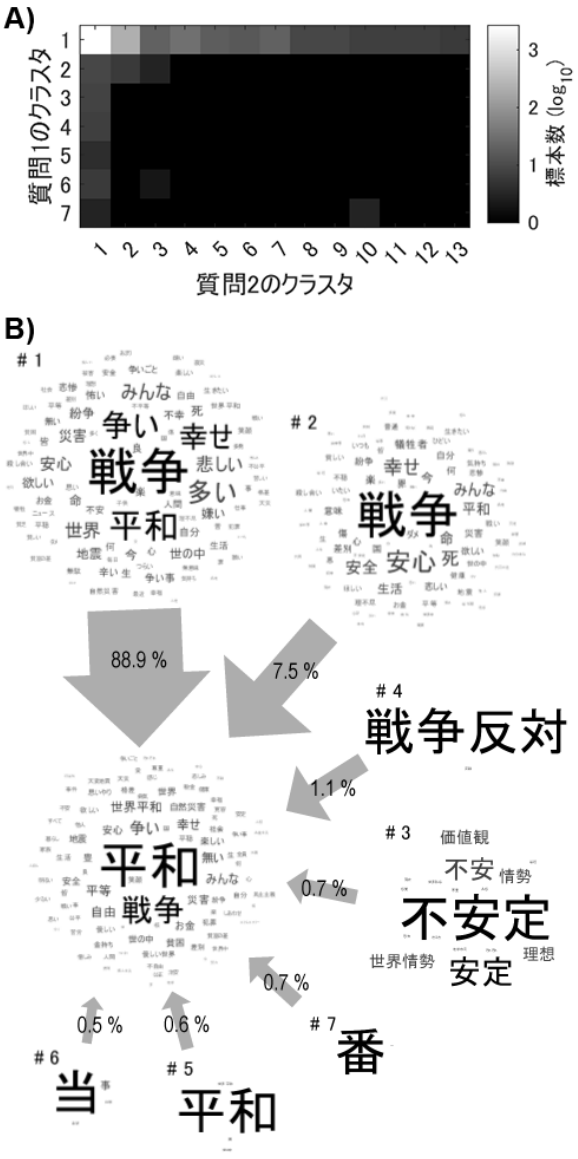


**図８クラスタ間の関係**

**A)** 推定された質問１と２のクラスタへ属する回答者数の二次元グレースケールマップ。縦軸方向に質問１から推定したクラスタ、横軸方向に質問２から推定したクラスタを示す。クラスタのインデックスは図７と対応する。属する標本数は対数軸で可視化した。

**B)** 質問１と２のクラスタの関係の例。質問１から推定したクラスタ #1 の理由に対応する質問２のクラスタ（A の図の最上列）のうち、10 人以上の回答者が属するものをワードクラウドで可視化し、質問１のクラスタ #1 のワードクラウドへの矢印を向けて配置した。#付の番号は質問 2 のクラスタのインデックスを表す。また、質問１のクラスタ #1 に属する回答者のうちで質問２の各クラスタに属する条件付き確率を、矢印の上にパーセント表記で示す。

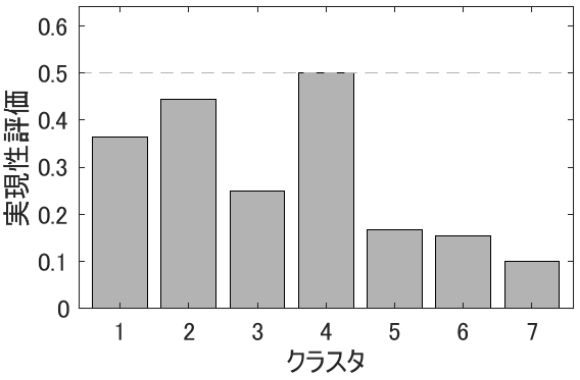


**図９．質問１のクラスタにおける実現性評価**

質問１から推定した理想の世界の回答クラスタにおける、質問３から求めた将来的な実現性に対するポジティブな評価の割合を示す。破線はチャンスレベルを表す。

性の高い分析であると言える。なお探索半径$\varepsilon$の値は、小さくすると意味空間上においてより近い語をグループ化することになり、グループの柔軟性を操作できる。今回は類似度行列（距離行列）の分布を参考に設定したが、語意に対する主観判断テストを実施するなどして、実際の語意判断の閾値の分布に基づいて設定する方策も有効だと考えられる。

模擬調査により得られたクラスタ（図７）では、関連する語を含む記述文をグループ化できた。しかし、直感とは一致しない結果も含まれている。質問１のクラスタでは漢字表記の「平和」と平仮名表記の「へいわ」が分かれた。両者の意味空間上のコサイン類似度は 0.498 であり、「平和」と「戦争」間の類似度の 0.581 よりも低い。これは学習元である Wikipedia において、平仮名表記が漢字表記とは異なる文脈で用いられることを反映している。距離の計算には回答文の語間の距離の中央値を用いており、他の語との組み合わせも影響したと考えられる。

質問 2 では多様な回答が得られ、多数のクラスタが推定される結果となった。しかしその多くは戦争に関連する回答であり、質問 1 と質問 2 のクラスタ同士の関係性については標本数の偏りが大きく（図８Ａ参照）、十分な検討ができなかった。また、質問１を受けた質問であるために、単体では意図を十分に判別できない回答も多かった（「当然」「一番（大事）」など）。提案手法のような語レベルに分解するクラスタリングの目的には、回答者に意図を十分に表現する語を回答させるような、独

立した質問文の設計が望ましいと考えられる。また単語の過剰な分解（「一番」→「一」「番」など）を避けるため、意味空間への埋め込みに用いる語の目的に応じた最適化方法についても検討する必要がある。

推定した回答クラスタの二次的な分析例（図９）では、質問１の回答クラスタごとに質問３の実現性評価の回答を分析し、クラスタの種類によって異なる評価傾向があることを示した。特に、「尊重」「平等」や「弱者」といった社会的な相対関係を示す語や「不安」「ストレス」といった内面的な語に対して実現性を低く評価しており、現代日本社会が抱える課題を反映するような結果が得られたことは興味深い。世論調査の分析に提案手法の枠組みを適用することで、様々な社会問題の可視化と政策評価の分析に寄与できることが期待される。なお本研究では標本数が確保できなかったため実施しなかったが、複数の質問回答のクラスタを組み合わせた分析も可能である。大規模な調査データを取得すれば、異なる視点の質問から回答者の分類精度を高めて世論の傾向を可視化できると期待される。

提案手法は世論調査への適用を念頭に、政治的関心事項の可視化というトピックの抽出に特化した手法となっている。Word2Vec は分布仮説に基づくため、対義語も近い距離に配置される性質がある。そのため、トピックに対する評価がポジティブなのか、ネガティブなのかといった意向の判別は多くの場合困難である。模擬調査への適用結果でも、前述の「戦争」と「平和」という対義語が同じグループに属する分類となっており、属する語の分布だけでは真意を読み取れない（図７）。政治的意向の可視化を目的とする場合、例えば政治的関心のグループ化を行ったのちに、属する文章を改めて自然言語処理によって文意の分析を行うなどの工夫が必要となる。

本研究では文章間の類似度に、それぞれの文章に含まれる語の間の類似度の中央値を用いた。例えば文章自体をベクトル化する Dec2Vec（Dai et al. 2015）のように文章の分類タスクに対して最適化することも可能だが、政治的関心事項の可視化という側面では語への分解のほうが最終的なクラスタの解釈面で優位である。一方で、文章中のひとつの語のみに着目した距離となるため、分類の頑健性には乏しいという課題がある。実際、特に回答が多様化した質問 2 では、有効回答中のクラスタに割り当てられた回答数の割合が半数を切っている。この比率はDBSCAN の密度閾値パラメータを小さくすることで減らすことは可能だが、クラスタの細分化を招くためにより距離の定義に依存した結果を導いてしまう。近年は前後のコンテクストの違いを考慮した単語の埋め込みが可能な深層学習モデルも提案されており（Devlin et al. 2018）、今後世論の可視化という目的により適したエンベディングや距離計算、クラスタリング手法についてシミュレーション等を行って検討していく必要がある。

近年、様々な分野において大量のデータを活用したボトムアップな分析方法が取り沙汰されている。例えば動物行動学の分野では、映像データから教師なし学習によってボトムアップに行動の単位へ分節化を行う計算論的エソロジー（computational ethology）と呼ばれるアプローチを採用した研究が報告されている（Anderson and Perona 2014）。データ自体に含まれるパターンを機械学習によって読み解くことで、従来の研究者がトップダウンに付与したラベルに基づく分析では見えてこない機序が明らかになると期待される。こうしたボトムアップな方法論は、動物行動学に限らず、研究者の主観的判断が介在するあらゆる科学研究分野に対して示唆を与えうるものである。調査研究ではこうしたアプローチに耐えうる標本数の収集が可能であり、世論や投票行動の分析に用いることで、従来は可視化されにくかった多様な世論について検討が可能になると期待できる。また、客観的な分類に基づく変数設定と定量化は計算論に基づく分析と親和性が高く、社会心理学や神経科学など、ヒトを対象とする周辺研究領域との学際的研究の可能性を広げることにも大きく寄与する。しかし一方で、ボトムアップな機械学習によって得られる結果は用いたアルゴリズムやパラメータに依存する側面があることに注意せねばならない。学習モデルのオープンソース化や生データの公開など、再検証可能な枠組みを構築し、再現性に対する多面的な検証を継続することで、現実的な機械学習を用いた新たな世論の可視化が可能になるだろう。

（埼玉大学 非常勤講師・慶應義塾大学 グローバルリサーチインスティテュート）

**謝辞**

本研究は科研費若手研究(23K16984)「非言語性社会信号を認識する脳内メカニズムの計算論的解明」による支援を受けた。

**参考文献**

Allahyari, M, Pouriyeh, S, Assefi, M, Safaei, S, Trippe, ED, Gutierrez, JB and Kochut, K (2017). A brief survey of text mining: classification, clustering and extraction techniques, *arXiv*, 1707.02919.

Anderson, DJ and Perona, P (2014). Toward a science of computational ethology, *Neuron*, 84, 18-31.

Bengio, Y, Ducharme, R and Vincent, P (2000). A neural probabilistic language model, *Advances in neural information processing systems*, 13, 1-7.

Dai, AM, Olah, C and Le, QV (2015). Document embedding with paragraph vectors, *arXiv*, 1507.07998.

Devlin, J, Chang, MW, Lee, K and Toutanova, K (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv*, 1810.04805.

Ester, M, Kriegel, HP, Sander, J and Xu, X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, *KDD '96: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 226-231.

平田 崇浩, 大隈 慎吾 (2021).「ノン・スポークン(Non-spoken)調査」の理念と課題, *政策と調査*, 20, 77-86.

Kudo, T (2005). Mecab: Yet another part-of-speech and morphological analyzer, http://mecab.sourceforge.net/.

Mikolov, T, Chen, K, Corrado, G and Dean, J (2013a). Efficient estimation of word representations in vector space, *arXiv*, 1301.3781.

Mikolov, T, Sutskever, I, Chen, K, Corrado, GS and Dean, J (2013b). Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, 26, 3111-3119.

三浦 麻子 (2020). 心理学研究法としてのウェブ調査, *基礎心理学研究*, 39, 123-131.

大栗 正彦 (2022). 参院選の報道各社の情勢調査比較 ー多様化する調査手法ー, *政策と調査*, 23, 5-22.

鈴木 正敏, 松田 耕史, 関根 聡, 岡崎 直観, 乾 健太郎 (2016). Wikipedia 記事に対する拡張固有表現ラベルの多重付与, *言語処理学会第 22 回年次大会発表論文集*, 797-800.