令和５年度学士学位論文

大規模言語モデルによる分散表現を用いた女性議員の活動状況の分析

八戸工業大学 工学部

システム情報工学科

瓜田 壮一郎

大規模言語モデルによる分散表現を用いた女性議員の活動状況の分析

G206012　瓜田 壮一郎

世界経済フォーラムは、各国の男女格差の現状を評価する「Global Gender Gap Report」の 2023 年版を発表した。我が国のジェンダーギャップ指数は過去最低の 146 ヵ国中の 125 位であり、特に政治参画の分野は 138 位と低い水準にある。政治参画の分野は、国会議員と閣僚の男女比、最近 50 年における行政府の長の在任年数により評価されており、実際の女性議員の議会等における活動状況は考慮されていない。

そこで本研究では、国会本議会における議員の発言に着目した。各発言に対して、大規模言語モデルを用いた分散表現の作成とクラスタリングによる発言のクラスタ形成を行い、議会において話されている話題を把握する。女性議員の人数の割合と議会での各話題に対する発言数の割合を比較することで、人数比だけでは捉えられない女性議員の活動状況の一面を把握することを試みた。

分析には、2018 年から 2022 年までの国会本会議における議員の発言を用いた。議員の 1 回の発言が複数の文からなる場合、その中で複数の話題に触れられている可能性がある。そのため、単一の話題が含まれるように発言を正規表現によって文単位に分割し、その１つ１つを発言として扱うこととした。それらの発言に対して BERT と OpenCalm-7B を用いて分散表現を作成した。その後、k-means 法によるクラスタリングを行うことで発言を話題別のクラスタに分割した。クラスタリングの結果を元に、各議員の国会本議会における各クラスタの発言割合と発言数の時系列変化を量的・視覚的に可視化した。そして、女性議員数の割合と全体の女性議員の発言数の割合及び話題別の女性議員の発言数の割合を比較することで、議会における女性議員の諸分野における活動状況を人数比に対する増加率で量的に把握した。

結果として、5 年間の国会本会議における発言全体を見た場合には、女性議員の人数比に対して発言数の割合が多くなっていた。また、クラスタ別に見ると発言数の割合にばらつきがあり、特に国内(地域)経済と質問の話題に関して、顕著に割合の低下が観察された。

Analysis of Female Legislators' Activities Using Distributed Representations by Large Language Models

G206012 Soichiro Urita

The World Economic Forum released the Global Gender Gap Report 2023, assessing the state of gender disparities across countries. Japan's ranking is the lowest ever, placed at 125th out of 146 countries, with the realm of Political Empowerment being notably low at 138th. The assessment of Political Empowerment is based on the gender ratio of parliamentarians and ministers, along with the tenure of executive heads over the past 50 years, not accounting for the actual activities of female legislators.

This paper, therefore, focuses on the talks made by legislators during the national parliament's plenary sessions. A distributed representation of each talk is generated using a large language model. And clustering to form clusters of talks to understand the topics being discussed in the Congress. Comparing the ratio of female legislators to the ratio of their talks on each topic attempts to uncover facets of female legislators' activities that are not reflected solely by the number of legislators.

This analysis utilized talks from the plenary sessions of the parliament from 2018 to 2022. Considering that a single talk by a legislator could include multiple sentences, potentially covering a variety of topics, these talks were segmented into sentence units utilizing regular expressions. Each segment was then considered an independent talk. For these segments, distributed representations were generated using BERT and OpenCalm-7B. Subsequent clustering through the k-means method divided the speeches into topic-specific clusters. Based on the clustering results, the study quantitatively and visually analyzed the proportion of talks per cluster and the temporal changes in the number of talks by each legislator. By comparing the ratio of female legislators to the overall proportion of talks made by female legislators, as well as the proportion of speeches by female legislators by topic, the research quantitatively assessed the activities of female legislators in various fields within the parliament, measured by the rate of increase relative to their numerical ratio.

The findings indicated that, over the five years of plenary sessions in the parliament, the proportion of talks made by female legislators was higher than the number of female legislators. Furthermore, when viewed by cluster, there was variability in the proportion of talks, with a significant decrease in the proportion of talks related to Japan’s economics and inquiries.

令和５年度学士学位論文

大規模言語モデルによる分散表現を用いた女性議員の活動状況の分析

八戸工業大学工学部

システム情報工学科

学籍番号　G206012　氏　名　瓜田　壮一郎

指導教員　島内　宏和

目次

第１章

序論

１．１　研究の背景

世界経済フォーラム(WEF)は、各国の男女格差の現状を評価した「Global Gender Gap Report」(世界男女格差報告書)の2023年版[1]を発表した。我が国のジェンダーギャップ指数は146カ国中125位となっており、2006年の公表開始以来、最低であった。特に政治分野は138位と低い水準にある。政治分野は、国会議員および閣僚の男女数の比、最近50年における行政府の長の在任年数の男女数の比により評価されており、実際の女性議員の議会等における活動状況は考慮されていない。

2010年代初頭から、政治分野でのテキストデータを用いた研究が進められている[2]。また、言語モデルを用いた例として、中国共産党のイデオロギーの長期的変化を把握するために、20年分の機関紙のテキストデータとDoc2Vecによりその傾向の変化を分析したもの[3]等がある。言語モデルについて、GoogleによりTransformerアーキテクチャを用いた言語モデルBERT[4]が提案されて以来、文書分類や固有表現抽出、類似文書検索などに広く活用されている。近年、OpenAIによるGPT-4[5]等をはじめとしたより大規模な言語モデルが注目を集めている。

そこで、議会における発言に着目し、大規模言語モデルを用いた分散表現への変換とクラスタリングによる話題の把握を行い、女性議員の人数の割合と発言数の割合を比較することで、人数比だけでは捉えられない女性議員の活動状況の一面を把握できるのではないかと考えた。

１．２　研究の方針

本研究では、国会本会議での発言を対象として、大規模言語モデルを用いた話題のクラスタ形成から議会で話されている内容を把握した上で、女性議員の発言割合を算出し、ジェンダーギャップ指数の評価指標である国会議員の男女比と比較することで、発言による女性議員の活躍を定量的に分析することを試みる。

１．３　本論文の構成

本論文の構成は以下のとおりである。第 1 章では、研究の背景及び方針について述べる。第 2 章では、準備として本研究に用いる技術の説明を行う。第 3 章では、発言内容のデータの抽出及びデータの前処理と抽出したデータの概要について記す。第 4 章では、使用する大規模言語モデルを選定し、大規模言語モデルを用いた分散表現の生成とクラスタリングによる話題の把握を行う。第 5 章では、クラスタリングの結果を用いて、議員別の発言の量的・質的要約を検討する。第 6 章では、女性議員の人数比と発言量の割合の比較から、国会本会議中における女性議員の議会への参画状況を把握することについて述べる。第 7 章では、それぞれの結果をまとめ、考察を行う。

第２章

準備

本研究に用いる主要な技術について説明を行う。２．１節では大規模言語モデルの説明と本研究で用いる言語モデルの説明を記す。２．２節では、クラスタリングとクラスタリング手法であるK-means法、その評価指標であるシルエット係数について述べる。２．３節では、使用したプログラミング言語とパッケージについて記す。

２．１　大規模言語モデル

大規模言語モデルとは、日本語版の Wikipedia や Common Crawl 等の大量のテキストデータを用いて事前学習を行った数十億から数千億のパラメータを超えるニューラルネットワークであり、文章生成や文章部類、固有表現抽出などタスクに用いることができる。

２．１．１　Transformer

Transformer[6]はRNNやCNNを用いず、全結合層とattention mechanismsに基づいたシンプルなエンコーダ・デコーダモデルである。RNNと異なり、並列処理を可能とすることで処理が高速化された。また、単語間の関連を直接モデル化することが可能となり、テキストの理解において前後の文脈を全体的に考慮することができる。

２．１．２　BERT

BERT（Bidirectional Encoder Representations from Transformers）[4]は、2018年にGoogle社によって提案された。Transformerアーキテクチャを利用して開発された大規模言語モデルであり、文脈に基づいた単語の表現を学習することで、自然言語処理タスクにおけるモデルの性能を大幅に向上させた。

２．１．３　OpenCalm-7B

OpenCalm-7B[7]は、サイバーエージェント社が開発した日本語に特化した商用利用可能な大規模言語モデルである。日本語版のWikipediaとCommon Crawlのオープンデータを使用して事前学習を行っている。GPT-NeoX[8]をベースモデルとしてトレーニングされ、デコーダのみのTransformer構造を採用している。

２．１．４　Calm2-7B-Chat

Calm2-7B-Chat[9]は、LLaMA[10]がベースモデルであるCalm2-7B[11]をチャット形式にチューニングした商用利用可能な言語モデル。既存のLLMをベースとした継続事前学習ではなく、１から新規に構築されたモデルであり、約1.3兆トークンの日本語と英語の公開データセットで事前学習している。

２．２　クラスタリング

本研究におけるクラスタリングは、大量のテキストデータから、同じような話題のグループ(クラスタ)を形成することである。

２．２．１　k-means法

k-means法は、データをk個のクラスタに分割するクラスタリング手法である。各クラスタ内のデータとそのクラスタのセントロイド(重心)との距離の総和を最小化することを目的として、k個のセントロイドをランダムに選択し、各データを最も近いセントロイドと同じクラスタに割り当てる。次に、割り当てられたデータの平均位置を計算し、セントロイドを更新する。この割り当てと更新を繰り返し、クラスタの割り当てに変更がなくなるまで実行する。

２．２．２　シルエット係数

シルエット係数は、クラスタリングの品質を評価するための指標である。クラスタ内のデータがどれだけ密接かを表す凝集度と、異なるクラスタとどれだけ離れているかを表す乖離度を用いて、値が-1から1まで範囲で算出される。1に近いほど適切にクラスタリングされている可能性が高いことを示す。

２．３　使用したプログラミング言語とパッケージ

後述するデータ収集やシステムの構築等においてプログラミング言語にPython(バージョン3.10.12)を用いた。Pythonは、ライブラリが豊富であり、機械学習や自然言語処理等の分野で広く使われている。コーディング環境にはJupyter Notebookを用いた。主要なパッケージとして、データハンドリングはPandas、ニューラルネットワークにはPyTorch、クラスタリングにはscikit-learn、グラフの作成にはMatplotlibを採用した。表2-1に使用したパッケージとそのバージョンを記す。

表2-1 使用したパッケージとバージョン

パッケージ名	バージョン
accelerate	0.26.1
ipywidgets	7.7.1
janome	0.5.0
japanize_matplotlib	1.1.3
matplotlib	3.7.1
numpy	1.23.5
pandas	1.5.3
scikit-learn	1.2.2
seaborn	0.13.1
torch	2.1.0
transfomers	4.35.2
wordcloud	1.9.3

第３章

データの収集

３．１節では、抽出対象の発言内容のデータを検討する。３．２節では、発言内容のデータの抽出方法について述べる。３．３節では、抽出したデータの前処理について確認する。３．４節では、最終的なデータの概要について述べる。

３．１　抽出対象の発言内容のデータ

本研究では、議会における各議員の発言に着目した。議会の中でも、衆議院・参議院の両院における主要な会議であると考えられる国会本会議を対象として、国会本会議における議員の発言内容のデータ収集し、分析を行うこととした。

３．２　発言内容のデータの抽出方法

発言内容のデータの抽出及びその発言者名等の収集方法には、国立国会図書館が提供している国会会議録検索システム検索用 API[12]を用いた。このシステムでは、国会における本会議や委員会での発言内容がデータとして記録されており、第 1 回国会(1947 年 5 月開会)分から保存されている。

国会会議録検索システム検索用 API を用いてデータを収集するにあたり、2018 年から 2023 年までの期間を対象とした。この理由として、データ収集時には 2023 年分の本会議がまだ終わっていなかったということに加え、「Global Gender Gap Report」の 2023 年版が 6 月に公開されており、それ以前の女性議員の発言率の変化について観察したかったため以下に記す 1 年間の発言の量と計算資源の状況を踏まえ 5 年間程度としたことが挙げられる。

３．３　抽出したデータの前処理

抽出した各議員の発言を見てみると、各議員の 1 回の発言が複数の文からなる場合があった。1 回の発言の中で複数の話題に触れられている可能性あると考えられたため、発言中にできるだけ単一の話題が含まれるように発言を正規表現によって文単位に分割し、その１つ１つを発言として扱うこととした。また、空白や記号、定型文などの重複する文については削除した。

３．４　データの概要

議事録から発言等を抽出し、前処理した後のデータの期間、議員数、全体及び各年のデータ件数を表 3-1 に示す。

表 3-1 抽出したデータの概要

期間	2018 年 ～ 2022 年
議員数	690 名
5 年間のデータ数	170,501 件
2018 年	43,426 件
2019 年	32,689 件
2020 年	24,449 件
2021 年	34,384 件
2022 年	35,553 件

データ件数について、議事録から抽出した段階では、データ件数が 228,714 件であり、前処理後は 170,501 件となった。重複文の削除をはじめとする前処理によってデータ件数の削減が行われたが、議員数の減少は見られなかった。表 3-1 で示したデータは、発言内容に加えて、発言者の名前・発言日・性別の情報からなるデータセットとなっている。

第 4 章

分散表現とクラスタリングを用いた話題の把握

大規模言語モデルにより発言内容の分散表現を獲得し、クラスタリングにより発言を話題毎にクラスタを形成する。４．１節では大規模言語モデルの選定について記す。４．２節では、分散表現の作成を行う。４．３節ではクラスタリングについて述べる。

４．１　大規模言語モデルの選定

本研究に用いる大規模言語モデルには、BERT(Bidirectional Encoder Representations from Transformers)[4]と、サイバーエージェントが開発した OpenCalm-7B[7]を選択した。

BERT は 2018 年に Google が発表したニューラル言語モデルである。Google Scholar での論文の引用件数は 90,165 件(2024 年 1 月 30 日 確認)であり、言語モデルとして広く利用されている。本研究では、東北大学の自然言語処理グループが公開している訓練済み日本語 BERT モデル[13]を用いた。

OpenCalm-7B は 2023 年 5 月 16 日にサイバーエージェントにより公開された商用利用可能な日本語に特化した大規模言語モデルである。事前学習として、日本語 Wikipedia と日本語版 Common Crawl が用いられている。

OpenAI が提唱したスケーリング則では、言語モデルのサイズ、データセットの大きさ、事前学習の計算量の増加に伴って、言語モデルの性能が向上するとされている。そこで、言語モデルとして広く利用されている BERT とパラメータ数が BERT の約 20 倍程度であり、事前学習に用いられたデータセットの種類が多い OpenCalm-7B を比較することとした。

４．２　分散表現の作成

選定した 2 つの大規模言語モデルを用いて、それぞれの発言内容を類似した発言が近い方向を向くようなベクトルである分散表現に変換する。変換された分散表現は言語モデルに依存する。また、今回用いる分散表現は、両言語モデルともネットワークより出力されたベクトルの平均とする。

４．３　クラスタリングによる文章のグループ化

生成した分散表現にクラスタリングを適用し、類似した発言を話題別にまとめる。本研究ではクラスタリングの手法に k-means 法を採用した。そのハイパーパラメータであるクラスタ数はシルエット係数を用いて決定する。クラスタ数が 2 から 29 の範囲でシルエット係数を算出し、最大値または極大値を与える点を候補とする。

BERT 及び OpneCalm-7B による分散表現に対して、クラスタ数を 2 から 29 に設定して k-means 法を適用した際の、クラスタのシルエット係数の推移を図 4-1 に示す。

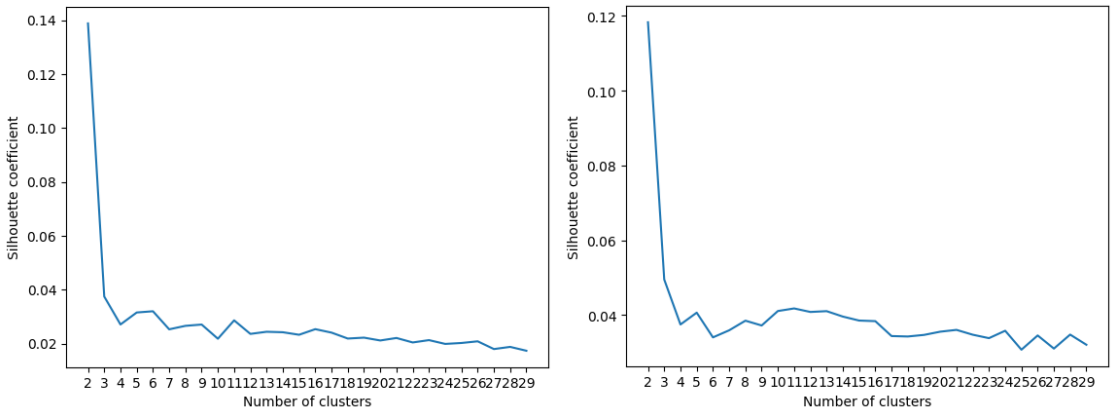

図 4-1　BERT(左)および OpenCalm-7B(右)による
クラスタのシルエット係数の推移

BERT ではクラスタ数 2 で最大、6 で最大の極大値をとっており、11 がそれに続いている。OpenCalm-7b の場合は 2 で最大、11 で最大の極大値で、5 で次に大きな極大値をとっている。国会本会議においては様々な話題が扱われることを考慮し、ここでは両者のクラスタ数として 11 を採用した。

各言語モデルの分散表現に対してクラスタスタリングを行った。また、各クラスタの内容を把握するために、クラスタ内の単語の頻度に基づくワードクラウドを作成した。このワードクラウドで表示される単語と、クラスタの中心に近い 20 件の発言をコサイン類似度により抽出した内容から、そのクラスタにおける話題を把握する。例としてワードクラウドと中心 5 件の発言を図 4-2 と表 4-1 に示す。

図 4-2 BERT(左)および OpenCalm-7B(右)による
財政に関するクラスタのワードクラウド

表 4-1 BERT(左)および OpenCalm-7B(右)による

クラスタの中心に近い 5 件の発言

政府の説明では、平成三十一年度の税収は、バブル期であった平成二年度の六十・一兆円を超えて、史上最高の六十二・五兆円になると説明しています	すなわち、今回の補正予算においても所得税は約四千五百億円の上方修正がなされており、国民の所得は引き続き上昇していることがうかがわれます
衆議院の審議で我が党の笠井亮議員が、ＪＯＧＭＥＣは二〇〇四年の設立以来、出資した案件のうち六割が生産に至らず事業を終結し、二〇一一年度に初めて繰越欠損金を出して以降、その額は、この十年で二十倍以上に当たる二千八百億円にも増大していることを明らかにしました	政府は、今般の窓口負担の引上げによる給付費減一千八百八十億円のうち、約半分の九百億円は一定の受診控えが起こるという、いわゆる長瀬効果として試算しているとのことです
安倍政権では、財政出動のための新規国債発行額は、二〇一二年度には四十七・五兆円でしたが、一九年度には三十二・六兆円と、約十五兆円も減少、これは、その分、財政出動の金額が抑制され続けたとも言えます、プライマリーバランスの黒字化をアピールするためだけに	また、臨時特別の措置と別に実施される一・一兆円規模の軽減税率も、財務省の試算によると、低所得層の軽減額が一千四百億円であるのに対して、高所得者層の軽減額は二倍以上の二千九百億円に上り、逆進性を助長することが明確です
この保育の受皿三十二万人分については、二十五歳から四十四歳までの女性の就業率が二〇二〇年度末に他の先進国並みの八割まで上昇することを想定して、必要な整備量を推計したものであります	他方、税収は、最近までの収入実績等を勘案して、約二兆三千二百億円の減収を見込んでおります
総理は、一月二十日の施政方針演説の中で、日本経済はこの七年間で一三％成長し、来年度予算の税収は過去最高となりました、公債発行は八年連続での減額であります、経済再生なくして財政健全化なし、この基本方針を堅持し、引き続き二〇二五年度のプライマリーバランス黒字化を目指します、この六年間で生産年齢人口が五百万人減少する一方で、雇用は三百八十万人増加しておりますなどと述べております	また、トリガー条項の発動が一年間続いた場合、地方で五千億円程度の減収が見込まれますが、地方の減収分を国費で補填するなど、地方財政の安定にも十分配慮するとともに、トリガー条項の効果の及ばない灯油や業界で燃料として使う重油の高騰対策も強化するよう求めます

上記のワードクラウドとクラスタの中心付近の発言の内容からクラスタの内容を把握し、それぞれにタイトルを付けた。各言語モデルにおける 11 個のワードクラウドとそのタイトルは付録 A に記した。

両言語モデルを確認したところ、共通する話題として国内外における経済政策、地域における政策、規制、質問、財政のクラスタが観察された。その他のクラスタについては話題の傾向が異なり、BERT では自己紹介からなる定型文に近いクラスタや、矛盾への指摘、責任追及などの話題がクラスタとして形成された。OpenCalm-7B においては、進行(定型)や政権批判が OpenCalm-7B でのみクラスタが形成された他、ワードクラウドとクラスタ中心の発言を確認しても内容が不鮮明なクラスタが 3 つあった。その中でも発言数が極端に少ないクラスタが 1 つ観察された。

第 5 章

議員別の発言の量的・質的要約

各議員の活動状況を量的・質的に把握するために、各議員の各クラスタの発言数の 5 年間における割合と、1 年ごと発言数およびクラスタの割合の時系列変化を可視化するシステムについて記す。５．１節では、可視化の手法と要約について述べる。５．２節では、作成した可視化システムの結果を記す。

５．１　可視化手法と要約

本研究で取り扱う国会本会議における各議員の活動状況を量的に把握するために、クラスタリングの結果を元に、各議員の各クラスタの発言数の 5 年間における割合を円グラフにより可視化し、1 年ごと発言数およびクラスタの割合の時系列変化は積み上げ棒グラフによって可視化を行う。また、議員の発言の要約には、サイバーエージェントによる対話用に指示学習が行われた calm2-7b-chat[9]を用いる。該当の言語モデルは、第 4 章において記した OpenCalm-7B の後続モデルであり、対話に特化している。

議員の発言と要約の指示をプロンプトとして、議員の 5 年間における発言の要約を生成・表示する。要約のプロンプトは「USER:次の文章を簡潔に要約してください。文章 :“(議員の発言)” ASSISTANT:」を用いた。

５．２　可視化システムの結果

議員を選択すると、対象議員の各クラスタにおける発言量の割合を示す円グラフ、各クラスタにおける発言量の時系列変化を表す積み上げ棒グラフ、議員の発言の要約文を表示するシステムをプログラミング言語にPythonを用いて構築した。該当システムによる円グラフ、積み上げ棒グラフ、要約文の一例を図5-1と図5-2、表5-1に示す。

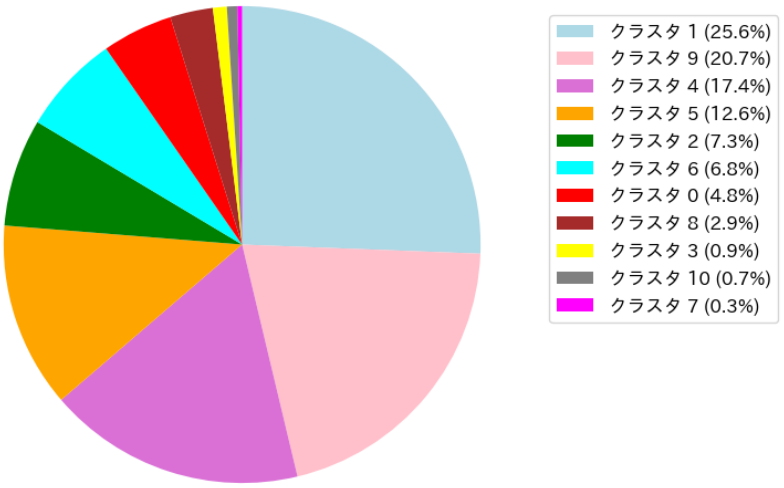

図 5-1 ある国会議員の各クラスタにおける発言比

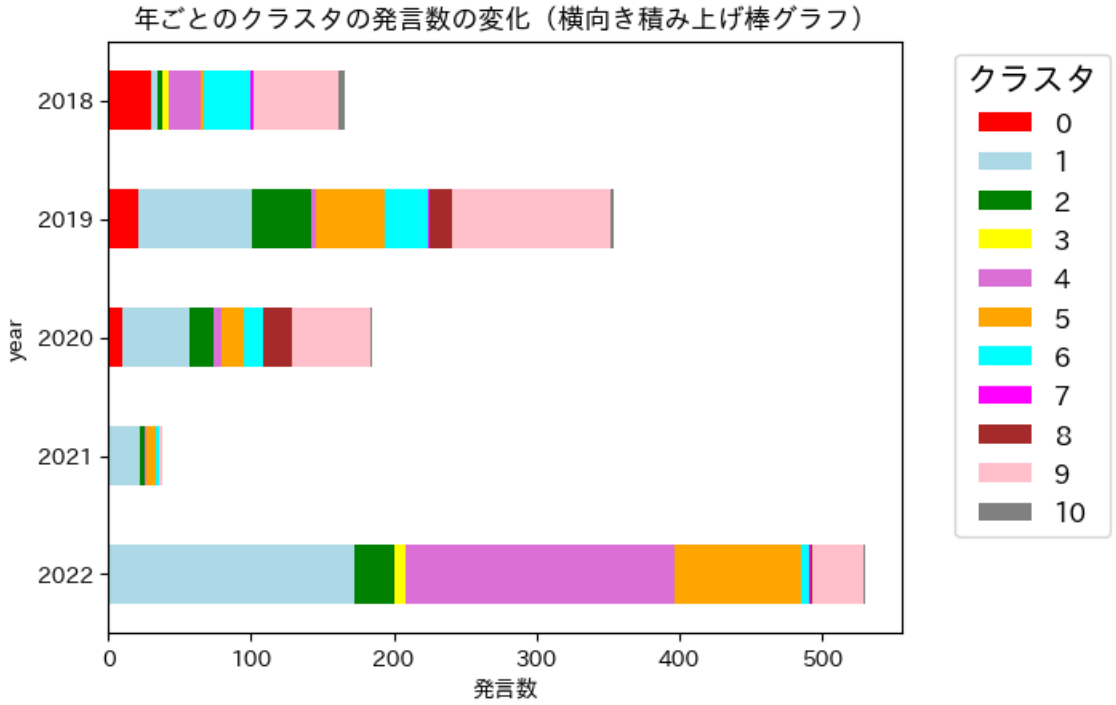

図 5-2 ある国会議員の発言数の時系列変化

表 5-1 ある国会議員の発言の要約

日本は、自由で開かれたインド太平洋を実現するために重要な役割を果たすべきです。また、外交努力の一環として、多くの国々の外相に直接提案し、理解を得ました。さらに、世界人口が年々増え、過去最多の約7000万人に達している現状に対して、日本は平和と安定に貢献することが求められています。また、中東地域の安定は日本を含む世界の安定に直接関係しているため、日本は中東の平和と安定に努めていきます。

5 人の国会議員の発言の要約を確認したところ、主観的には該当政治家の公約と一致しているように見受けられたが、本研究では要約に対して定性的な評価しかできていない。また、同様の発言データおよびプロンプトを用いて複数回生成した結果、極端に短い文章や「ありがとうございます」などの定型文、データ中で頻出度の高い単語のみが表示される場合があり、生成される要約の内容に大きなばらつきがあった。

第 6 章

女性議員における人数比と発言量の割合の比較

女性議員の人数の割合とは発言量の差を量的に把握するために、クラスタリングによる各話題の発言数を用いて女性の発言割合を算出し、国会議員の男女比と比較をする。６．１節では、比較に用いる女性議員の人数比と発言量の割合の概要を述べる。６．２節では、比較を行う過程と結果について記す。

６．１　比較に用いる女性議員の人数比と発言量の割合

女性議員の人数の割合については、本研究で使用する 2018 年から 2022 年の内、2018 年から 2021 年までを男女共同参画局が、男女共同参画社会基本法に基づき作成している年次報告書である男女共同参画白書[14]を参考として、各年度における衆議院・参議院における女性議員の割合を収集した。各年における両院の女性議員の割合と 5 年間全体の女性議員の割合を表 6-1 に示す。

表 6-1　女性議員の人数の割合

	衆議院	参議院
定員	465 名	248 名
2018 年	10.1%	20.7%
2019 年	10.2%	20.7%
2020 年	9.9%	22.9%
2021 年	9.9%	23.0%
2022 年	9.9%	25.8%
全体	14.3%	

収集したデータの女性議員の割合から、5 年間全体での女性議員の割合は約 14.3%であった。

６．２　女性議員の全体における発言量とクラスタ別の発言量の割合

収集したデータ数から女性議員が発言したデータ数を抽出し、女性議員の全体における発言量の割合を算出した。また、各クラスタ内の女性議員の発言量は、そのクラスタ内での全体のデータ数から女性議員が発言したデータ数を抽出し、算出した。女性議員の全体における発言量の割合を表 6-2、言語モデル別の各クラスタ内の女性議員の発言量を表 6-3 に示す。

表 6-2　女性議員の全体における発言量の割合

全体のデータ数	170,501 件
女性議員の割合	15.3%

表 6-3　両言語モデルによる各クラスタの女性議員の発言量の割合

BERT		OpenCalm-7B	
自己紹介	22.9%	不鮮明・少量	20.0%
矛盾への指摘	20.5%	進行	19.4%
責任追及	19.7%	地域政策	18.4%
厚生	18.1%	不鮮明	18.4%
回答	18.1%	政権批判	18.1%
対外経済政策	17.6%	不鮮明	18.1%
財政	17.1%	規制	17.1%
国民の声	13.3%	財政	15.6%
規制	13.1%	質問	12.6%
国内経済政策	8.4%	地域経済	9.5%
質問	4.6%	対外政策	6.5%

表 6-1 で示した女性議員数の割合と表 6-2、表 6-3 で示した発言数の割合を比較し、女性議員数の割合に対する発言数の割合の増加率を算出する。

女性議員数の割合に対する女性議員の全体における発言量の割合の増加率を表 5-5 に示し、女性議員数の割合に対する両言語モデルによる各クラ

スタの女性議員の発言量の割合の増加率を表 6-6 に示す。また、各クラスタにおける発言の割合の増加率について計算した結果の記述統計量を表 6-7 に示す。

表 6-5　人数の割合に対する全体における発言量の割合の増加率

全体のデータ数	170,501 件
全体の女性議員数の割合	14.3%
女性議員の発言の割合	15.3%
増加率	+7%

表 6-6　人数の割合に対する両言語モデルにおける
各クラスタの女性議員の発言量の割合の増加率

BERT		OpenCalm-7B	
自己紹介	+60%	不鮮明・少量	+40%
矛盾への指摘	+43%	進行	+36%
責任追及	+38%	地域政策	+29%
厚生	+27%	不鮮明	+29%
回答	+27%	政権批判	+27%
対外経済政策	+23%	不鮮明	+27%
財政	+20%	規制	+20%
国民の声	−7%	財政	+9%
規制	−8%	質問	−12%
国内経済政策	−41%	地域経済	−34%
質問	−68%	対外政策	−55%

表 6-7　言語モデル別の各クラスタの増加率の記述統計

	BERT	OpenCalm-7B
最小値	-68%	-55%
第１四分位数	- 8%	- 1%
第２四分位数	+23%	+27%
第３四分位数	+32%	+29%
最大値	+60%	+40%

表 6-5 から、女性議員数の割合に対する 5 年間全体の女性議員の発言数の割合の増加率は+7%となっていた。5 年間全体で議員数の割合に比べて発言率の割合の方が大きかった。また、表 6-6 から、両言語モデルとも 11 クラスタの内７つのクラスタにおいて女性の発言割合が 20%以上増加していることが観察され、話題別に比較した際にも過半数のクラスタで議員数の割合に比べて発言率の割合が大きかった。両言語モデルにおいて共通して見られた話題の内、財政(BERT:+20%, OpenCalm-7B:+9%)、質問(BERT:-68%, OpenCalm-7B:-12%)、国内(地域)経済(BERT:-41%, OpenCalm-7B:-34%)のクラスタが両言語モデルで増加・減少の傾向が一致していた。

表 6-7 において、中央値である第 2 四分位数の値は BERT が+23%、OpenCalm-7B が+27%であり、四分位数からも過半数の話題において、議員数の割合に比べて発言率の割合が大きかったと言える。また、最大値は BERT が+60%、OpenCalm-7B が+40%であり、最小値は BERT が-68%、OpenCalm-7B が-55%であることから、クラスタ別の発言数の割合における増加率は散らばりがあることが観察された。

第 7 章

考察

第 4 章のクラスタリングにおいて、形成されるクラスタは言語モデルに依存する点に注意が必要である。BERT と OpenCalm-7B を用いて抽出された各 11 のクラスタの内、国内外への政策や質問、規制等の話題からなるものが共通して観察された。しかし、それ以外のクラスタについては傾向が異なった。また、OpenCalm-7b によるクラスタの中に、発言数が極端に少なく、内容にばらつきのあるものが存在した。この要因として、k-means 法においてクラスタが超球状に分布していることを仮定していることなども考えられる。

第 5 章で作成したシステムを用いることで、各議員の国会本議会における各クラスタの発言割合と発言数の時系列変化を量的・視覚的に把握することができた。要約において、生成される要約文の質にばらつきがあることと定性的な評価しか出来ていないことが課題として挙げられる。

第 6 章の比較について、2018 年から 2022 年までの 5 年間における国会本会議全体では、女性議員数の割合に対して女性議員の発言量が 7%高くなっていた。クラスタごとの増加率の四分位数と最大値・最小値を踏まえると両モデルともクラスタ別の男女比にばらつきがあり、女性の発言率が顕著に少ないものもあった。例として、国内経済、質問のクラスタがある。共通するクラスタの傾向は一致しており、該当クラスタの結果の頑健性が示唆される。

第８章

結論

5 年間全体における女性議員の発言と過半数の話題における発言の割合が女性議員の人数の割合より高く、国会本会議においては人数の割合以上に活動が活発であることが示された。政治分野におけるジェンダーギャップ指数の評価指標と国会本会議内での発言量に基づく女性議員の活動状況に乖離がある可能性が示唆される。増加率の最も高い話題が BERT における自己紹介のクラスタであることから、女性議員の発言機会が比較的多く与えられていることが示唆される。一方、増加率が顕著に低い話題として両言語モデルで共通している国内(地域)経済と質問のクラスタがある。議題として国内(地域)経済のみが共通して発言率が低いことから、5 年間の国会本会議においては女性の経済分野への参画が限定されていると考える。また、質問の発言率が低いことについては、全体的に発言数が人数の割合に比べて多い一方で、積極的に女性が質問できる環境ができていない可能性がある。女性議員の質問が少ないことで政治的議論における多様性の欠如をもたらし、政策を決定する上で女性の視点が十分に反映されない可能性がある。各クラスタにおいて顕著に発言率が高い話題と低い話題があることから、人数に基づく評価だけでは女性議員の詳細な活動を完全に捉えることが出来ないため、ジェンダーギャップ指数における評価指標の見直しにも検討の余地があると考える。

今後の課題として、本研究と同様の分析方法で他国におけるジェンダーギャップ指数の評価指標との差異を分析し、国別の比較をすることができると考えられる。異なる政治・文化的背景を持つ国々におけるジェンダーギャップの実態と、それに対する政府の対応に関する洞察を深めることができる可能性があると考える。分析方法における課題として、内容が不鮮明な話題がクラスタリングされないようにすることと、他の大規模言語モ

デルでも共通となるクラスタを特定することでクラスタリングされる話題の頑健性を高めることが挙げられる。そのために、大規模言語モデルに対して、議事録を用いてファインチューニングすることと、よりパラメータ数の多いものを使用することを検討する。また、他のクラスタリングの手法についても検討し、適切なモデルを決定するとともに結果についても精査していきたいと考えている。

参考文献

[1] Global Gender Gap Report 2023, https://jp.weforum.org/publications/global-gender-gap-report-2023/,(参照 2024-02-06)

[2] 福元健太郎. (2020). 政治学における人工知能の応用へ向けて. 人工知能, 35(4), 526-533.

[3] 御器谷裕樹, 持橋大地. (2022). 文書ベクトルを用いた中国共産党のイデオロギーの分析. Group, 2(2), 4.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] GPT-4, https://openai.com/research/gpt-4, (参照 2024-02-06)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[7] OpenCalm-7B, https://huggingface.co/cyberagent/open-calm-7b, (参照 2024-02-06)

[8] GPT-NeoX, https://huggingface.co/docs/transformers/model_doc/gpt_neox, (参照 2024-02-06)

[9] calm2-7b-chat, https://huggingface.co/cyberagent/calm2-7b-chat, (参照 2024-02-06)

[10] LLaMA,
https://huggingface.co/docs/transformers/model_doc/llama,
(参照 2024-02-06)
[11] Calm2-7B,
https://huggingface.co/cyberagent/calm2-7b,
(参照 2024-02-06)
[12] 国会会議録検索システム,
https://kokkai.ndl.go.jp/#/, (参照 2024-02-06)
[13] berth-base-Japanese-whole-word-masking
https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking, (参照 2024-02-06)
[14] 男女共同参画白書,
https://www.gender.go.jp/about_danjo/whitepaper/index.html,(参照 2024-02-06)

謝辞

本研究を進めるにあたり、御指導と御鞭撻を頂いた島内宏和先生に深く感謝申し上げます。

そして、ともに研究を行った菊池真緒氏、松倉巧氏、松江海音氏、三田知広氏、三橋令旺氏、谷地村暢也氏、並びに八戸工業大学の方々に対し感謝申し上げます。

付録A

BERTとOpenCalm-7Bを用いた時の話題のワードクラウドとそのタイトル

両言語モデルを用いて作成された分散表現により、クラスタリングを行った結果のワードクラウドと、そのタイトルを以下に示す。

自己紹介

国民の声

対外経済政策

質問

財政

責任追及

図 A-1　BERT における話題のワードクラウド(1)

規制

国内経済政策

矛盾の指摘

厚生

図 A-2　BERT における話題のワードクラウド(2)

政権批判

地域経済

財政

図 A-3　OpneCalm-7B における話題のワードクラウド(1)

不鮮明、少量

規制

質問

不鮮明

不鮮明

対外政策

地域経済

進行

図 A-4　OpneCalm-7B における話題のワードクラウド(2)