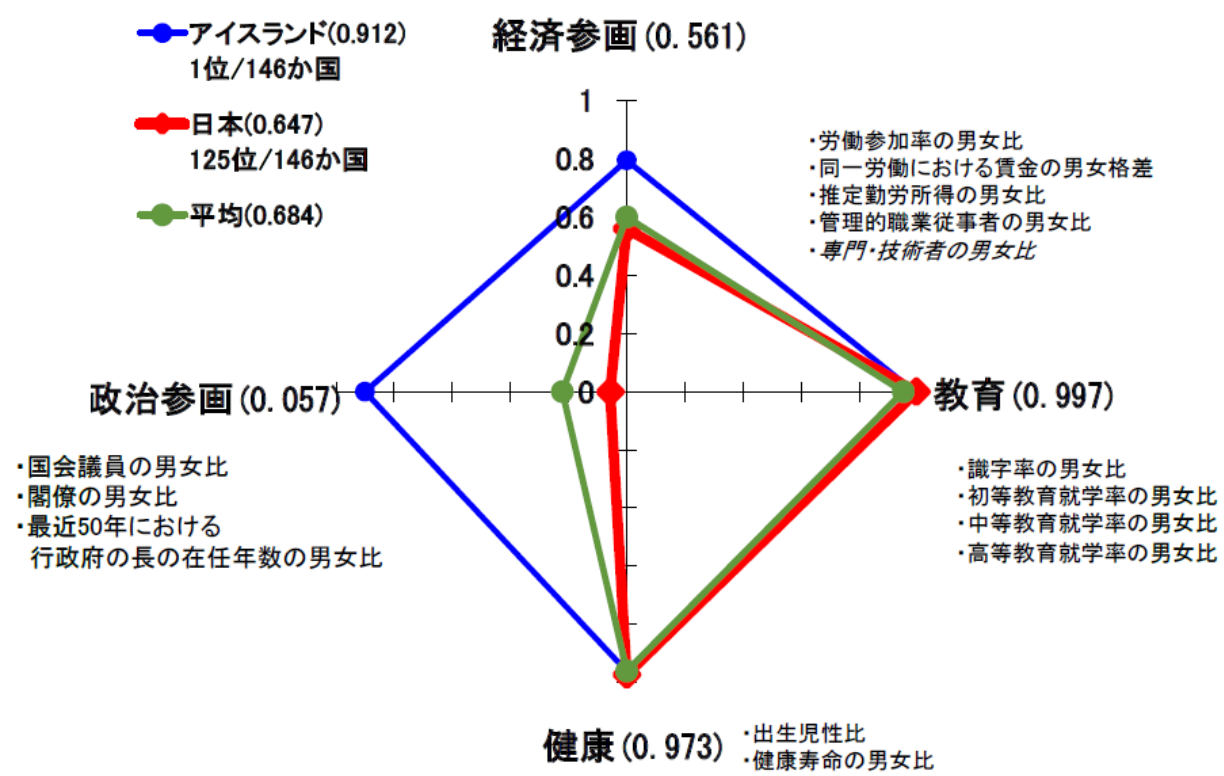
大規模言語モデルによる分散表現を用いた女性議員の活動状況の分析

瓜田 壮一郎(G206012)
八戸工業大学 工学部 システム情報工学科
卒業研究発表会 8-103　　2024年2月1日(木)

背景

ジェンダー・ギャップ指数（GGI） 2023年

・スイスの非営利財団「世界経済フォーラム」が公表。男性に対する女性の割合（女性の数値/男性の数値）を示しており、**0が完全不平等、1が完全平等**。
・**日本は146か国中125位。「教育」と「健康」の値は世界トップクラスだが、「政治」と「経済」の値が低い。**



（備考）1. 世界経済フォーラム「グローバル・ジェンダー・ギャップ報告書(2023)」より作成
2. 日本の数値がカウントされていない項目はイタリックで記載
3. 分野別の順位：**経済（123位）**、教育（47位）、健康（59位）、**政治（138位）**

順位	国名	値
1	アイスランド	0.912
2	ノルウェー	0.879
3	フィンランド	0.863
4	ニュージーランド	0.856
5	スウェーデン	0.815
6	ドイツ	0.815
15	英国	0.792
30	カナダ	0.770
40	フランス	0.756
43	アメリカ	0.748
79	イタリア	0.705
102	マレーシア	0.682
105	韓国	0.680
107	中国	0.678
124	モルディブ	0.649
125	日本	0.647
126	ヨルダン	0.646
127	インド	0.643

特に政治参画の分野では
138 位/146ヵ国中

公開開始以来**最低**

日本の順位は
125 位/146ヵ国中

内閣府.男女共同参画に関する国際的な指数. https://www.gender.go.jp/international/int_syogaikoku/int_shihyo/index.html

目的

政治参画の分野における評価指標

- **国会議員の男女比**
- **閣僚の男女比**
- **最近50年における行政府の長の在任年数の男女比**

⇩

実際の女性の議会等における**活動状況**は考慮されていない

⇩

発言を大規模言語モデルとクラスタリングにより抽出し、
議会における**女性の発言率**と**国会議員の男女比**の比較から
活動状況の一面を量的に把握することを目指す

大規模言語モデルとは

OpenAI

事前学習として、大量のテキストデータを自己教師あり学習している数億から数千億のパラメータを持つ言語モデル

- Googleにより**BERT**が提案され、**文書分類**や**固有表現抽出**などに広く活用されている。
- OpenAIによるGPT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。
  - サイバーエージェントの**OpenCalm-7B**、Calm2-7b-chat等は詳細な仕様とともに公開されている。

手順１．データ収集

① APIを用いて、データを収集

② 発言が複数の文からなる場合、発言を文単位に分割

③ 空白や記号、定型文などの重複する文については削除

2018年～2022年 / 690名 / 170,501件

議事録

APIで抽出

各発言

正規表現で前処理

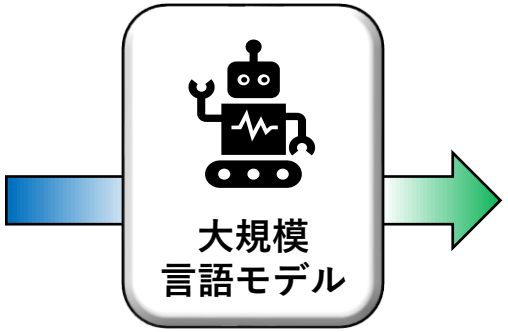
各議員の発言データセット（例）

日付	発言者名	発言内容
⋮	⋮	⋮
2019-05-16	本多平直	第四の問題点は、イージス・アショア、Ｆ35などの米国製高額兵器導入の犠牲となり、本当に必要な…
⋮	⋮	⋮
2021-04-01	鈴木淳司	このことを重ねて申し上げ、良識ある衆議院の皆様に対し、このような決議案を断固として否決してい…
⋮	⋮	⋮
2022-11-22	岸田文雄	物価高騰の要因については、基本的にはエネルギー、食料品を中心とした物価高であり、こうした分野…
⋮	⋮	⋮

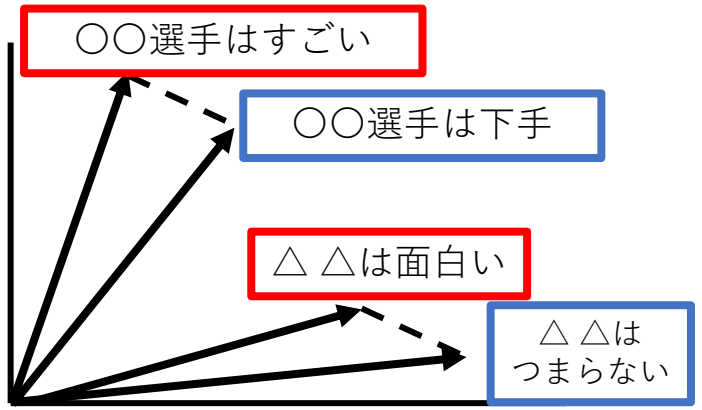
国立国会図書館.国会会議録検索システム.https://kokkai.ndl.go.jp/#/

手順２．分散表現

各議員の発言



各議員の発言の分散表現



分散表現とは

- 単語や文章の意味や関連性を数値ベクトルで表現する方法。
- **似た意味を持つ言葉は近く、関連性の低い言葉は遠く**に配置される。
- 生成される分散表現は言語モデルに依存する。

BERT

- Transformerを用いた言語モデルであり、広く利用されている
  - 提案した論文のGoogle Scholarでの引用数は、89,766件

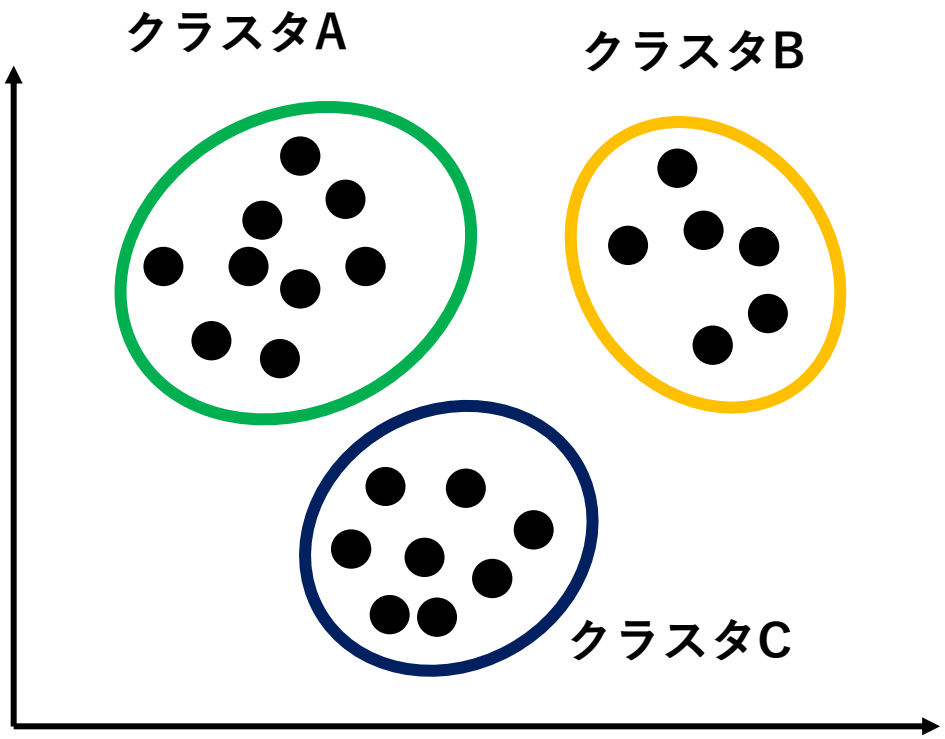
OpenCalm-7b

- パラメータ数がBERTの20倍
- 日本語に特化した言語モデル

OpenAIが提唱する言語モデルのスケーリング則によると
モデル及びデータセットのサイズ、計算量が多いほど、より高い性能を持つとされる

手順３．クラスタリング

分散表現をクラスタリングし、**話題**（似た内容）ごとにまとめる。



k-means法

- 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。
  - クラスタの中心からそのクラスタ内の各データまでの距離の総和が最小になるように最適化される。

シルエット係数

- クラスタ内の凝集度と、クラスタ間の乖離度を用いて、-1から1までの値を取る。
- シルエット係数が１に近いほど、凝集度・乖離度が高く良いクラスタ数といえる。

手順４．比較

５年間全体における女性議員の発言率

全体の発言数	170,501 件
女性議員の発言率	**15.3 %**

or

各話題における女性議員の発言率

比較 ⟷

女性議員の割合

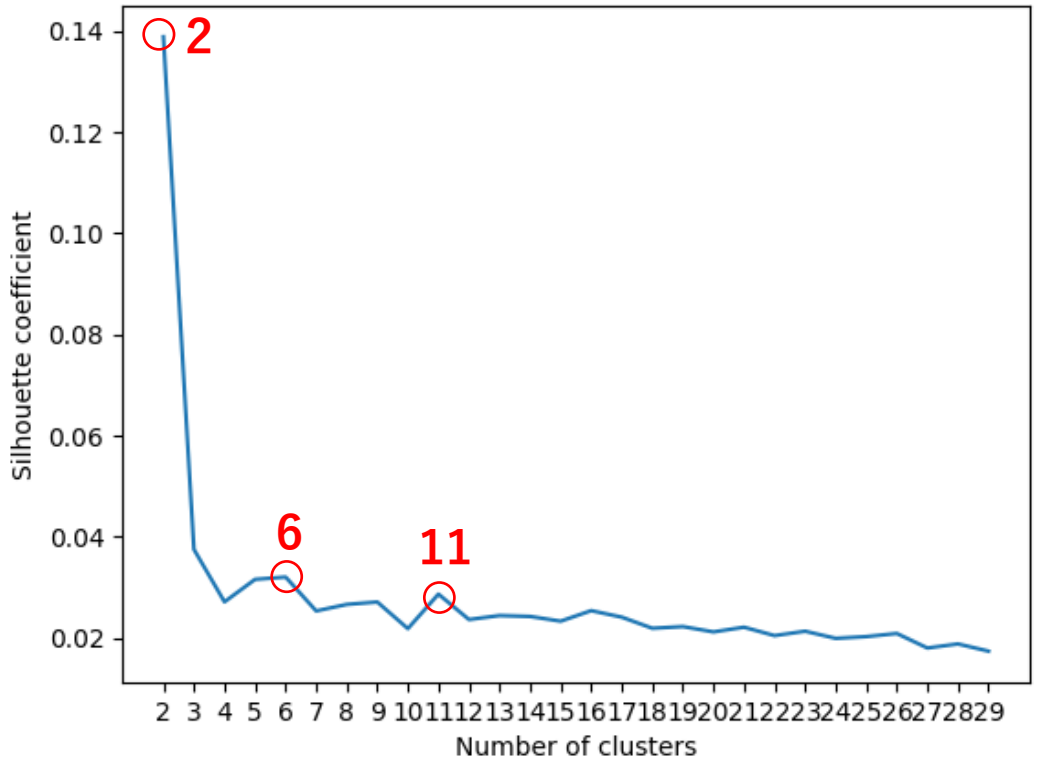
	衆議院	参議院
2018年	10.1 %	20.7 %
2019年	10.2 %	20.7 %
2020年	9.9 %	22.9 %
2021年	9.9 %	23.0 %
2022年	9.9 %	25.8 %
5年間	**14.3 %**	

$$\frac{\text{発言率}}{\text{人数の割合}} - 1 \rightarrow \pm X\%$$

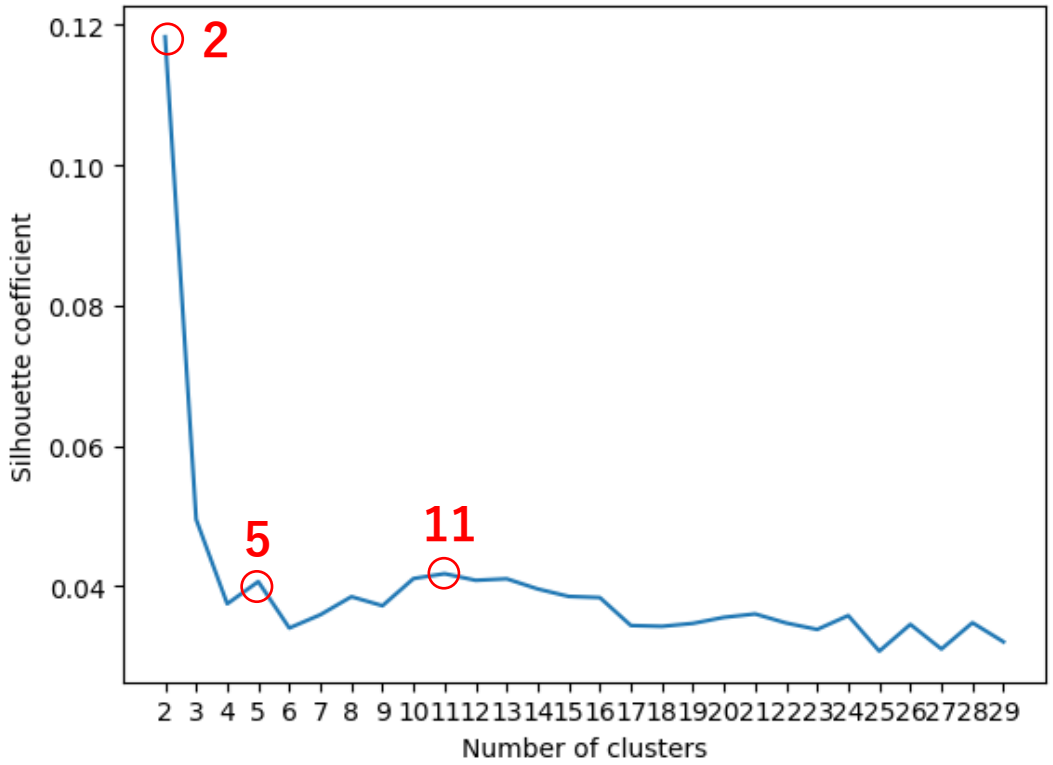
男女共同参画局.男女共同参画白書.https://www.gender.go.jp/about_danjo/whitepaper/index.html

結果 シルエット係数の推移

BERT



OpenCalm-7B



	最大値	極大値①	極大値②
BERT	2	6	11
OpenClam-7B	2	11	5

様々な話題を扱われることを考慮し
クラスタ数「11」を採用

結果 発言の割合と人数比の比較

$$\frac{\text{全体の発言率}}{\text{全体の人数の割合}} - 1 \rightarrow +7\%$$

人数比に対して発言率が、「+」であれば高く，「ー」であれば低いと示唆される。

BERT	
自己紹介	+ 60%
矛盾への指摘	+ 43%
責任追及	+ 38%
厚生	+ 27%
回答	+ 27%
対外経済政策	+ 23%
財政	+ 20%
国民の声	- 7 %
規制	- 8 %
国内経済政策	- 41%
質問	- 68%

OpenCalm-7B	
不鮮明・少量	+ 40%
進行	+ 36%
地域政策	+ 29%
不鮮明	+ 29%
政権批判	+ 27%
不鮮明	+ 27%
規制	+ 20%
財政	+ 9 %
質問	- 12%
地域経済	- 34%
対外政策	- 55%

まとめと考察

＜比較の結果について＞

- 5年間の本会議全体では、人数に対して発言量の比率が**大きかった。**
  - 各クラスタの**増加率には散らばり**があった。
  - 中には女性議員の**発言率が顕著に低い話題**もあった。
    - [例] 国内経済・質問のクラスタ
- **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。

＜クラスタリングについて＞

- 形成するクラスタは**言語モデルやクラスタリング手法に依存**している。
  - OpenCalm-7Bにおいて、発言数が極端に少なく、**内容が不鮮明なクラスタが存在**した。K-means法の仮定の影響なども考えられる。

- 他の言語モデルやクラスタリング手法とその結果について精査したい。

ご清聴ありがとうございました。