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

島内研究室　　瓜田 壮一郎（G206012）

１．はじめに　「Global Gender Gap Report」において、政治分野における日本の順位は公開開始以来最低であった。評価指標は、国会議員や閣僚の男女比等の人数を基本としたもので構成されており、実際の議員の活動状況は考慮されていない。そこで、本研究では大規模言語モデルを用いて、女性議員の活動状況の一面を議事録から把握することを検討する。

２．研究方法　2018 年から 2022 年における本会議で議論されている話題を大規模言語モデルとクラスタリングにより抽出した上で、各話題における女性の発言割合を算出し、国会議員の男女比と比較する。手順を以下に示す。

手順１．議事録から議員の発言を抽出し、発言内容や性別からなるデータセットを作成する。
手順２．それぞれの発言を大規模言語モデルにより分散表現に変換する。類似した発言内容が近い方向を向くベクトルに変換される。
手順３．分散表現に対し、k-means 法によるクラスタリングを適用し、似た内容の発言をまとめることで議会における話題を抽出する。
手順４．女性議員の 5 年間全体の割合および各クラスタ中における発言数の割合を比較する。

大規模言語モデルとしては、BERT が広く利用されている。実際、Google Scalar の引用数は 89,000 件以上となっている。他方、OpenAI の大規模言語モデルのスケーリング則によると、学習データ、計算量、モデルのサイズが大きいほど高い性能であるとされる。ここでは BERT と、パラメータ数が BERT の約 20 倍であり、日本語に特化している OpenClam-7B[1]の２つを用い、両者を比較しながら分析を進める。

３．結果　クラスタ数を変化させたときのシルエット係数のグラフの極値に着目し、両者で 11 をクラスタ数に採用した。両モデルで形成されたクラスタには、国内外への政策等の話題からなるものが共通して観察された。また、片方のモデルにのみ含まれるクラスタも存在した。

女性議員数の割合に対する、女性議員の発言数の割合の増加率は 7%となっていた。同様に、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 1 に示す。両言語モデルとも過半数のクラスタにおいて女性の発言割合が 23%以上増加している。共通して見られた財政・質問・国内経済については、両言語モデルで増加・減少の傾向が一致した。

表１　言語モデル別の各クラスタの増加率の記述統計

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

４．考察　5 年間の本会議全体では、人数に対して発言量の女性比率が多くなっていた。また、クラスタごとの増加率の四分位数と最大・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が、形成されるクラスタは言語モデルやクラスタリング手法に依存する。今後は、他のモデルも検証し、結果について精査したい。

５．参考文献(2024 年 1 月 24 日最終確認)

[1] OpenCalm-7B,　https://huggingface.co/cyberagent/open-calm-7b