大規模言語モデルによる分散表現を用いた国会議員の発言の量的・質的な要約の検討

瓜田　壮一郎*，島内　宏和(八戸工業大学)

Preparation of Papers for Student Workshop of IEEE IM Japan Chapter

Soichiro Urita, Hirokazu Shimauchi (Hachinohe Institute of Technology)

１．はじめに

世界経済フォーラムは、各国の男女格差の現状を評価したジェンダーギャップ指数の 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 カ国中 125 位となっており、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。そこで、本研究では議会における女性の活動状況の一面を量的・質的に把握することを目指し、その足掛かりとして議会における各国会議員の発言を量的・質的に要約するシステムについて検討する。

２．関連研究

社会科学の分野においてデータや機械学習を用いた研究が進められており、例えば政治学における研究が(1)に纏められている。言語モデルを用いたものとして、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(2)等がある。

言語モデルに関し、Google により Transformer アーキテクチャを用いた言語モデル BERT 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[1]。近年、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[2]。GPT は実装等の詳細な仕様は公開されていないが、日本語に特化した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。

３．研究方法

本研究では国会の本会議に焦点を当てる。議事録における議員の発言を大規模言語モデルによりベクトル化し、クラスタリングにより発言をクラスタに分け、議員の各クラスタにおける発言数の割合や時系列変化を可視化するシステムを構築する。また、指示学習により対話用にチューニングされた大規模言語モデルを用いた議員の全発言の質的な要約文も同時に表示する。システムの構築の詳細な手順は以下の通りである。

3.1 データの収集 2018 年から 2022 年までの国会本会議の議事録における 17,501 件の発言を分析対象とした。国立国会図書館による国会会議録検索システム検索用 API[3]を用いて、当該会議における各議員の氏名、発言内容、日付のデータを収集した。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場合がある。そのため、ここではできるだけ単一の話題が含まれるよう、発言を正規表現により文単位に分割し、その一つ一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。

3.2 分散表現の生成 各発言を言語モデルにより分散表現と呼ばれる、類似した発言が近い方向を向くようなベクトルに変換する。生成される分散表現は言語モデルに依存する。本研究では、先行研究において広く利用されている東北大学自然言語処理グループが公開した BERT[4]と、パラメータ数が BERT の約 20 倍程度であるサイバーエージェントが公開した OpenCalm-7b[5]の二つを用い、両者の結果を比較する。なお、分散表現は、両者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。

3.3 クラスタリング 生成した分散表現にクラスタリングを適用し、類似した発言をクラスタにまとめる。ここではクラスタリングの手法として k-means 法を採用し、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。また、クラスタの内容を把握するために、各クラスタ内における単語の出現頻度に基づくワードクラウドを作成する。

3.4 議員ごとの発言の量的・質的要約 議員の各クラスタの発言数の 5 年間における割合と、1 年ごとの発言数およびクラスタ割合の時系列変化を、それぞれ円グラフと積み上げ棒グラフで可視化する。また、議員の発言の質的な要約には、サイバーエージェントによる対話用に指示学習

[1] BERT を提案した論文の執筆時における Google Scholar での引用数は、83,542 件となっている。

[2] OpenAI が提唱する言語モデルのスケーリング則によれば、モデルおよ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。

[3] https://kokkai.ndl.go.jp/api.html

[4] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[5] https://huggingface.co/cyberagent/open-calm-7b

が行われた calm2-7b-chat[6]を用いる。議員の発言全体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。

4．結果

本章では、3 章に示した方法により形成した発言のクラスタと、クラスタの割り当てに基づく議員の発言の量的・質的な要約の結果を以下に示す。

4.1 クラスタリングにより抽出された話題の比較 BERT 及び OpenCalm-7b による分散表現に対し、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 1 に示す。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。抽出されたクラスタの例として、両者の予算に関するワードクラウドを図 2 に示す。

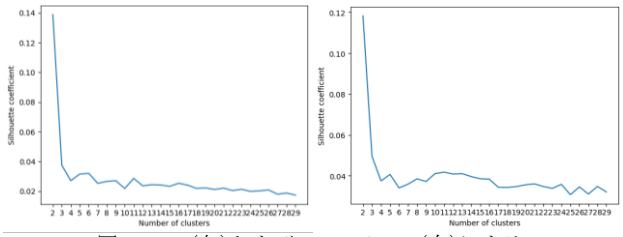

図 1 BERT(左)および OpenCalm-7B(右)による
クラスタのシルエット係数の推移

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

両者の各クラスタを確認したところ、国内外における経済政策、地域における政策、規制などの話題からなるクラスタや自己紹介などの定型文に近いものからなるクラスタ等が観察された。両者に共通しているクラスタとしては規制(BERT 10.8%, OpenCalm-7b 9.8%)、予算(BERT 6.1%, OpenCalm-7b 7.8%)、質問(BERT 4.9%, OpenCalm-7b 8.7%)に関するものが観察された。BERT についてはクラスタの割合は 1.2%～6.3%となっていた。OpenCalm-7b によるクラスタについては、発言数の割合が 0.3%と相対的に小さくなっているクラスタが 1 つ存在した。さらに、該当クラスタ内の発言の内容にはばらつきが見られた。

[6] https://huggingface.co/cyberagent/calm2-7b-chat

4.2 議員の発言の量的・質的要約 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 3 と図 4、表 3 に示す。

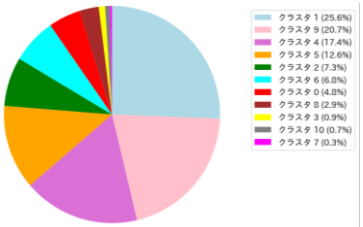

図 3 ある国会議員の各クラスタにおける発言比

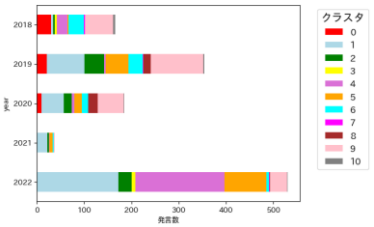

図 4 ある国会議員の発言数の時系列変化

表 3 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．考察

構築したシステムを用いることで、議員の議会における各クラスタの発言割合等を把握することができるが、形成されるクラスタは言語モデルにより異なる点には注意が必要である。2 つの言語モデルで抽出された 11 のクラスタのうち、3 つのクラスタは同種のものであったが、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が、k-means 法の仮定の影響なども考えられる。質的要約に関し、要約に用いたプロンプトでは要約文の文字数の制限等の指定はしていないが、より多くの会議や長期間にわたる議事録を扱う場合には対応が必要になる可能性もある。

6．まとめと今後の課題

本研究では、大規模言語モデルによる議事録上の発言の分散表現をクラスタリングし、各議員の発言を量的・質的に要約するシステムを構築した。今後は他の大規模言語モデルやクラスタリングの手法の変更についても検討を行った上で採用するモデルを決定し、議会における女性の活動状況の一面の分析に本システムを応用したい。

文　献

(1) 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.

(2) 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.