国会会議録を利用した議員の発言特性の抽出

有田智也　松井くにお
金沢工業大学工学研究科情報工学専攻
C6000578@planet.kanzawa-it.ac.jp

概要　本論文では Web 上に公開されている国会会議録を利用し，TF-IDF を用いた議員ごとの発言特性の抽出及び TF-IDF 値を用いた議員の関係性分析と，機械学習を用いて議員の発言を評価した結果を述べる．機械学習では議員の応答文を学習データとして seq2seq で学習させることで応答文を作成し，生成された応答文を利用し議員の発言の一貫性の分析を試みた．議員の発言をトークン分割する際には，Sentence Piece を利用することで，複合語を考慮しつつ特徴語の抽出ができる．また少数のデータを seq2seq で学習する際には，トークン分割に Sentence Piece を用いることで，トークン分割に MeCab を利用した場合に比べて，より自然で文法的に正しい応答文の作成が可能になり，高品質な関係分析が可能になる．

1 はじめに

近年，Apple 社の「Siri」や，Google 社の「Google Assistant」などの，知能的な対話システムの研究が活発化している．また女子高校生を模した，日本マイクロソフト社の「りんな」や，NTT 先端技術総合研究所が精華町役場と共同で研究している「京町セイカ」など対話システムに，ある特定の個性を持たせようとする研究も活発化している．しかしながら個性的な文章にするために，どの単語が役に立つのか，という研究は積極的になされていない．そこで本研究では個性付与の役に立つ言葉の抽出方法に着目した．

特徴語を抽出する研究として木村ら[1]は TF-IDF を用いて各地方議員の発言から，議員が重点を置いている政策を抽出しているが，抽出対象は名詞のみであり，助動詞や助詞などの議員の発言特性を抽出できそうな品詞を利用していない．また大南ら[2]は最大エントロピー法を用いて短命議員と長期議員に特徴的な単語を抽出し，”っぽい”などの短い個性を表す語の抽出には成功しているが，”であります”といった複数の単語から構成される意味を持つ語(複合語)を抽出することは出来ていない．そこで本研究では複合語での特徴語の抽出，抽出した特徴語での関係性の分析，及び複合語でトークン分割した文章を学習させたモデルでの応答文生成の評価を試みた．

2 利用した技術とデータ

・TF-IDF

TF-IDF(TermFrequency-InverseDocument Frequency)とはKaren Sparck Jones(1972)によって提唱された重みづけ手法であり，2種類の重みづけ手法を組み合わせることで実現されている．TFでは「ある文章中において出現回数の多い語はより重要な単語」という考え方に基づいて重みづけを試みる．IDFでは「ほかの文章中での出現回数が少ないほど重要な単語」という考え方に基づいて重みづけを試みる．つまりTF-IDF の値が大きければ大きいほどその文章にとって重要な語である．

・SentencePiece

SentencePiece[3]とは 2018 年に工藤拓らにより提案された入力文章をサブワードにトークン分割する手法である．SentecePiece では辞書を利用せず，コーパスから教師無し学習で文章をトークン分割するためのモデルを生成し，文章のトークン分割を行える．SentencePiece で国会会議録を分割した結果の一部を表 1 に示す．

表 1．sentencePiece での分割結果

原文	技能実習制度は今既にあり機能しているわけであります
トークナイズ後	技能実習制度/は/今/既に/あり/機能/しているわけであります
原文	考え方については先ほど法務大臣から答弁させていただいたとおりでございます
トークナイズ後	考え方/については/先ほど/法務大臣/から/答弁/させていただいた/とおりでございます
原文	みずから襟を正して職責を全うしてもらいたいと考えております
トークナイズ後	みずから/襟/を/正/して/職/責/を/全/う/してもらいたい/と考えております

・Seq2Seq

Seq2Seq(Sequence to Sequence)[4]とは 2014 年に Ilya Sutskever らによって提案された，入力された系列を別の系列に変換するモデルである．Seq2seq は Encoder と Decoder の二つの RNN から構成され，Encoder では入力系列を読み込み隠れ層の最後の状態を Decoder に渡す．Decoder では Encoder から与えられた隠れ層をもとに，入力出力終了を表す”<eos>”が現れるまで単語を出力する．RNN の部分を LSTM[5]や GRU[6]にすることで離れた語の依存関

係を学習させることが可能になる．

・国会会議録

話者が明示されている上，文法的な正しさが保証されているデータとして国会会議録を使用した．データの収集には国立国会図書館が公開しているAPI[7]を用いて，2012 年から 2019 年までの国会会議録を収集した．その後国会会議録から安倍晋三議員，麻生太郎議員，稲田朋美議員，福島哲郎議員，福島みずほ議員，蓮舫議員の発言を抽出した．また対話を学習させるためのデータには各議員の発言の内第一文までを利用した．

3 提案手法

図 1 に提案システムの構成を示す．国会会議録コーパスから特定の議員の発言内容を抽出し，発言データには発言内容をそのまま保存し，対話データには発言内容に加えて他の議員の質問内容を保存する．トークン分割には MeCab[8]と Sentence Piece を用いた．発言データでは TF-IDF の値を用いて議員ごとの特徴語を抽出する．対話データでは seq2seq を用いて議員ごとの発言内容を学習させる．

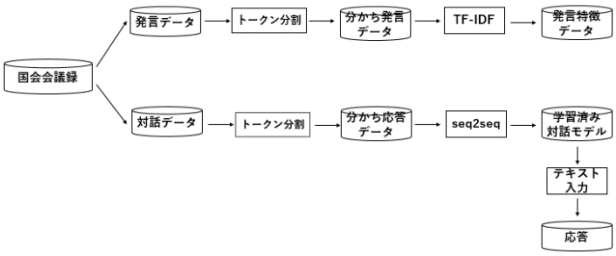

図 1.　提案システムの構成

3.1 TF-IDF を用いた特徴抽出と評価

2012 年 1 月 1 日から 2019 年 12 月 31 日までの安倍晋三議員の発言を MeCab[8]+neologd 辞書でトークン分割した後に TF-IDF の値が 0.1 以上の単語を抽出した結果を図 2 に示す．

言わば
いわば
あるいは
方々
つき
皆様
さまざま
しかし
デフレ
そうした
安倍政権
社会保障
におきまして

図 2.　MeCab で分割したデータでの特徴語

データを MeCab+neologd 辞書でトークン分割した場合”わけであります”や”だと思っております”などの複合語は，”わけ/で/あり/ます”や，”だ/と/思っ/て/おり/ます”のように細かく分割され,意味を持つ複合語から単体では意味の伝わらない単位にまで分割されてしまう．

この問題を解決するために本研究では SentencePiece を用いてトークン分割を試みた. SentencePiece を用いてトークン分割することで，高頻度で現れる語をそのまま利用することが可能になる. そのため”わけであります”などの高頻度で文章中に現れる語を SentencePiece でトークン分割した場合，複合語である”わけであります”のまま出力される．SentecePiece を用いて安倍晋三議員の文章を分割した後に TF-IDF の値を計算して，TF-IDF の値が 0.1 以上の特徴語を抽出した結果を図 3 に示す．

わけでございます
いわば
わけであります
言わば
わけでありまして
わけでございまして
であります
議員にお答えをいたします
こう
わけでありますが
あるいは
このように思います
ところでございますが
でございますが

図 3.　SentencePiece で分割したデータでの特徴語

トークン分割に SentencePiece を用いることで議員の発言の複合語も議員の特徴語として抽出することが可能になった．SentencePiece と TF-IDF を用いて 2012~2015 年の議員の発言の特徴語を抽出した結果を表 2 に示す．

表 2.　2012~2015 年の特徴語

安倍晋三	麻生太郎	稲田朋美	福山哲郎	蓮舫	福島みずほ
わけでございます	と存じます	改革	ありがとうございます	事業	社民党の
言わば	参々	公務員制度改革	おはようございます	ありがとうございます	福島みずほ
わけでございました	につきましては	法人	んですけど	参会	参芥
このように思います	金融	内閣人事局	属は	独法	お聞き
わけであります	御存じのように	基本法	個は	あるいは	雇用
わけでございますが	そういった	独法	言われた	機構	参考人に
わけでありますが	なんだと思います	幹部	御配慮賜	ありがとうございました	をいたします
わけでありまして	というもの	人材	参芥	従来型の	よろしくお願いします
であります	いかね	私経	御経歴会の	参芥	厚生労働省
いわば	と思っております	加新	いやいや	理事長	ありがとうございます
中において	なねね	というふうに	をいただいた	レビュー	生活保護
議員にお答えをいたします	案件	人事	秘密	武器	あるいは
あるいは	というものを	委員が御指摘	明確大臣	独立行政法人	労働者
このように	国税	まさしく	もう一回	契約	医療
しっかりと	記憶	きちんと	を含めて	んですけれども	参考人
我々は	んだと思います	ポスト	けど	公害	というふうに
であります	おられる	核軍	ありがとございました	行革	きちっと
上において	基本的に	国家	聴きます	校目	国際的自衛隊の行使
ただいま	基本的には	戦略	お答えください	201	介護
でございますが	意味で	お答え申し上げま	語に	ファンド	私目
このように思っております	おる	人事院	質案	自衛隊	消費者
デフレ	デフレ	国家公務員	関連書	医療	公述人
他が国	少なくとも	運転	特定秘密	実際委員	反発
につきましては	きちんと	主務大臣	いただけますか	研究	サービス
委員がご指摘	私どもとしては	というふうに思い	んですけれども	見直し	秘密
であります	おります	いたしております	個別的自衛権	戦略	いかがでしょうか
の中において	御存じのとおり	というふうに考え	ことについては	年金	ということでよろしいですね
においても	金融機関の	御指摘	実際委員	にお伺いします	すべき
んだろう	日本の	お答ね	お聞いたしたいと思います	教えてください	やる
でございまして	なんだ	自成	よろしくお願いしたいと思います	森村	だって

2016~2019 年の議員の特徴語を抽出した結果を表 3 に示す．黄色で示された部分は 2012~2015 年と 2016~2019 年で共通の特徴語である．

表 3.　2016~2019 年の特徴語

安倍晋三	麻生太郎	稲田朋美	福山哲郎	蓮舫	福島みずほ
わけでございます	恐々	ということでございます	じゃ	んですが	社民党の
いわば	私どもとしては	日報	実は	立憲民主党	福島みずほ
わけであります	につきましては	南スーダン	んですけど	じゃ	お聞き
言わば	そういった	でございます	お答えください	つまり	をいたします
わけでございますが	というもの	訓練	んですね	あるいは	厚生労働省
わけでありまして	御存じのように	紛争	でございます	統計学園	参考人に
わけでございまして	いろいろ	部隊	おはようございます	文科省	んですが
であります	というものを	派遣	けど	んですけれども	年金
誠意にお答えをいたします	おる	弾道ミサイル	んですが	んですね	労働
こう	いかぬ	我が国	いいですか	野川	時間ですので
わけでありますが	と思っております	である	外務大臣	安倍総理	あるいは
あるいは	ですけれども	委員がご指摘	分からない	憲法	過労死
このように思います	私どもとしては	pko	国は	教育委員会	賃金
ところでございますが	余裕	まさしく	近畿財務局	学生	んですね
でございますが	に関しましては	米軍	さっき	民進党の	駄目
中において	私どもは	自衛隊の	あった	実は	消費者
ということでございます	なんだと思います	承知をいたしておりますが	だって	お会い	厚労省
ところでございますが	と存じます	監督	立憲民主党	なんですが	じゃ
をさせていただいた	ならぬ	ところでございます	組織的犯罪集団	んでしょうか	というふうに思います
我々は	んですが	または	何で	過労死	会社
でありまして	という話	海幕	いやいや	二十	おかしい
ただいま	というのを	開示	ちゃんと	分からない	終わります
当然	んだと思います	御指摘	まあ	総理は	労働時間
につきましては	なん	済み	答えてください	ヒアリング	というふうに思います
でございます	おります	現行憲	にお伺いします	である	というか
でございまして	きちんと	に対する	言われている	メモ	ということでよろしいですね
こう思う	だと思っております	統幕	よろしくお願いしたい	分かりました	障碍者
である	というものは	米側	国交省	国家戦略特区	分からない
んですが	結構で	日米	財務省	投資	法案
に対する	国税	陸上自衛隊	というか	に伺います	労働能量性

次に単語の 2012~2015 年と 2016~2019 年での特徴語の一致率を表 4 に示す．

表 4.　特徴語の一致率

所属政党	議員名	一致率
自由民主党	安倍晋三	0.63
自由民主党	麻生太郎	0.5
自由民主党	稲田朋美	0.1
立憲民主党	福山哲郎	0.23
立憲民主党	蓮舫	0.03
社会民主党	福島みずほ	0.3

2012~2015 年と 2016 年~2019 年の特徴語一致率を比較すると，与党議員の安倍晋三議員と麻生太郎議員の特徴語一致率は 0.63，0.5 と高く特徴語が変化していない．これは質問に答える際に深く言及することを避け，抽象的な内容で返答したもしくは，多くの質問に答えているため，特徴語として特定の名詞が現れにくかったためである．しかし稲田朋美議員は与党議員にも関わらず，特徴語の一致率が 0.1 である．これは 2016 年に公務員精度改革を担当する国務大臣の職を退き，防衛大臣の職に就き，答える質問の内容が変化したためである．また野党議員の福山哲郎議員，蓮舫議員，福島みずほ議員の特徴語一致率は 0.23, 0.03, 0.3 と小さく，特徴語が大きく変化していた．これはその時ホットな話題について集中して質問することが多く，その質問した事柄が特徴語として現れたためである．

3.2 Seq2Seq を用いた応答文生成と評価

2012 年 1 月 1 日から 2019 年 12 月 31 日までの安倍晋三議員の答弁 25,459 件を MeCab+neologd 辞書で分割した後に他議員からの質問を入力，安倍晋三議員の応答を出力として Seq2Seq で学習させ，学習済みモデルを作成した．本研究で用いた Seq2Seq の中間層は 256，単語埋め込み次元は 300 である．学習させたモデルに対して，議員の発言からランダムに選んだものを学習済みモデルに入力し，生成された応答文を図 4 に示す．

入力	応答
全く納得いきません	先ほど申し上げましたようにこれはまさにこれはまさにこれはまさに
時間ですのでまとめて下さい	現在の交渉の中においてこれはまさにこれはまさにこれはまさに
ちょっと待って	現在の交渉の中においてこれは日本の国益を確保するために必要
お聞きしたいと思います	個別具体的な問題であります
よろしいでしょうか	当然のことであります
それはいつごろでしょうか	そうができません
静粛におねがいします	緩めていきたいと考えています

図 4.　MeCab で分割したデータを Seq2Seq で学習させたモデルから得られた応答文

MeCab で分割し，先頭の単語をランダムにサンプリングし，以降の単語を貪欲法でサンプリングすることで生成した応答文の問題は次の 2 つである．

・複合語が細かく分割されるため，学習することが難しく，複合語が出力の途中で崩壊してしまう．

・出力の内容が”これはまさにこれはまさに”のようにループし文章として成立しない．

複合語をそのまま利用し，複合語生成の学習を省略させるため，SentecePiece を用いて対話データを分割し，Seq2Seq で学習させた結果を図 5 に示す．

入力	応答
全く納得いきません	去年の三月の閉会審査において既に今治市の運営については猪瀬元秘書官が説明させていただきました
時間ですのでまとめて下さい	ホームページロンキーンで議論する招致をしたところでございます
ちょっと待って	確認をいたしました
お聞きしたいと思います	厳しい受け止めて言われると何度もところわかりやすく役割を担ってと認識しております
よろしいでしょうか	申し上げられないと思いますが
それはいつごろでしょうか	当時の十九万人につきましては明確にできません
静粛におねがいします	立場にはないことでございますのでそのことは受け止めておきたいとこのように思います

図 5.　SentencePiece で分割したデータを Seq2Seq で学習させたモデルから得られた応答文

対話データを SentencePiece でトークン分割し学習させることで，MeCab でトークン分割した場合，17.7%で発生していた応答文の崩壊と応答文のループを，6.7%に減少させることに成功した．

学習させたモデルに，議員の発言をランダムに 30 件抽出したものを入力し，生成された応答文に対して，文法的に正しい応答が生成される確率，自然な応答文が生成される確率，生成された応答文に TF-IDF 上位 30 件の単語がマッチする確率を評価した．

2012~2015 年の評価結果を表 5 に示す．

表 5.　2012~2015 年の応答文の評価

所属政党	議員名	文法的な正しさ	応答文の自然さ	TF-IDF上位30件
自由民主党	安倍晋三	55%	40%	35%
自由民主党	麻生太郎	50%	35%	35%
自由民主党	稲田朋美	45%	30%	20%
日本民主党	福山哲郎	50%	25%	25%
日本民主党	蓮舫	50%	35%	20%
社会民主党	福島みずほ	40%	5%	50%

2012 年~2015 年の評価の結果，自然な応答文が生成される確率は，与党議員に比べて野党議員の方が平均で 13%低かった．これは野党議員の発言内容は様々な分野への質問が多く，学習が困難であるのに対して，与党議員の発言内容は回答一辺倒であり，内容に一貫性が生まれ学習しやすかったためである．また応答文に TF-IDF 上位 30 件の単語が含まれる確率は与党議員に比べて福島瑞穂議員を除く野党議員の方が平均で 8%低かった．これは与党議員の特徴語に複合語が多く，複合語の方が名詞に比べて応答文に反映させる余地が多いためである．福島みずほ議員の応答文に含まれる単語が TF-IDF 上位 30 件にマッチする確率が高いのは”~の福島みずほです”という応答文が頻繁に生成されたためである．

2016~2019 年の評価結果を表 6 に示す．

表 6.　2016~2019 年の応答文の評価

所属政党	議員名	文法的な正しさ	応答文の自然さ	TF-IDF上位30件
自由民省	安倍晋三	55%	50%	35%
自由民主党	麻生太郎	55%	40%	40%
自由民主党	稲田朋美	45%	30%	10%
立憲民主党	福山哲郎	40%	20%	30%
立憲民主党	蓮舫	35%	20%	5%
社会民主党	福島みずほ	50%	20%	5%

2012~2015 年と 2016~2019 年を比較すると蓮舫議員の応答文の文法的な正しさ，応答文の自然さ，応答文が TF-IDF 上位 30 件にマッチする確率が減少している．これは 2012~2015 年に比べて 2016~2019 年の方が，より様々な政治的事柄について質問するようになったため，学習が難しくなったためである．

4 おわりに

本論文では，国会会議録を用いて議員の発言特性の分析と抽出を TF-ID を用いた手法と対話モデルを作成する手法で行った．SentencePiece を用いることで複合語の抽出が可能になり，高品質な個性抽出，応答文生成ができる．

参考文献

1．地方議会会議録コーパスプロジェクト(引用日:2010 年 12 月 20 日)http://local-politics.jp/

2．大南勝，掛谷秀紀．国会会議録に基づく短命大臣の特徴分析第 2 報．言語処理学会第 23 回年次大会論文集，pp．1167-1170，2018

3．Taku Kudo，John Richardson．SentencePiece:A simple and language independent sub word tokenizer and tokenizer for Neural Text Processing.
https://arxiv.org/pdf/180806226.pdf

4．Ilya Sutskever, Oriol Vinyals, Quoc V. Le．Sequence to Sequence Learning with Neural Networks
https://arxiv.org/abs/1409.3215

5．Sepp Hochreiter, Jurgen Schmidhuber．Long Short-term Memory.
https://www.researchgate.net/publication/13853244_Long_Short-term_Memory

6．Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio.
Learning Phrase Representations using RNN Encoder-Deocder for Statistical Machine Translation.
https://arxiv.org/abs/1406.1078

7．国会会議録検索システム検索 API(引用日:2020 年 12 月 20 日)
https://kokkai.ndl.go.jp/api.html

8．工藤拓，山本薫，松本裕治．Conditional Random Fields を用いた日本語形態素解析．Applying Conditional RandomFields to Japanese Morphological Analysis．https://ci.nii.ac.jp/naid/110002911717