地方議会議事録にもとづく議会発言特徴の抽出

大住 恭平 [1]　名取 良太 [1]　竹中 要一 [1,2]
[1] 関西大学 総合情報学部　[2] 大阪大学 大学院医学系研究科
k616355@kansai-u.ac.jp　takenaka@kansai-u.ac.jp
natori@res.kutc.kansai-u.ac.jp

1 はじめに

地方議会議事録とは、各地方議会がその議会における発言等を記録したものであり、各地方議会によって公開されている。

しかし現在、地方議会議事録に関する解析は TF-IDF を用いた特徴語抽出などはされているが、その他の手法を用いた解析はあまりされていない [1]。

そこで本研究では地方議会議事録を自然言語処理や機械学習などの様々な手法を用い解析を行うことで地方議会の発言から特徴抽出を行う。自然言語処理では日本語評価極性辞書を用い発言にネガポジスコアを付与する。

次にスコアに基づき発言者の性別や所属といった属性の推定を機械学習に行う。属性の推定を高精度に行うことができれば、それは地方議会において議員は性別や所属といった属性に応じた特徴をもつ発言を行なっている事を意味する。この試みが、女性議員の増加や所属政党の多様性を保つことによって議論に幅が広がることを定量的に評価する方法論として役立つこと期待している。

2 解析に用いるデータ

2.1 地方議会議事録データベース

地方議会議事録データベースとは、関西大学総合情報学部名取ゼミで開発, 管理されている、地方議会における議事録などから発言や議案、議員情報などを、定量データとして格納したものである [2]。

本研究では本データベースに格納されている 2011 年から 2015 年の大阪府の全市町村の地方議会のものを用い解析を行う。

2.2 日本語評価極性辞書

日本語評価極性辞書とは東北大学乾・鈴木研究室が公開している辞書であり、日本語の単語に対してネガティブ, ポジテイブなどの情報を付与したものである [3]。

日本語評価極性辞書には用言編 [4] と名詞編 [5] があり、それぞれネガポジの与え方が異なる。そのため本研究ではそれらを [-1,1] の数値で表現する。用言編では、ネガ=-1, ポジ=1 とし、名詞編では、n=-1,e=0,p=1 とした。

3 解析手法

3.1 日本語評価辞書を用いたネガポジ分析

ネガポジ分析としてデータベース内の全発言に対し日本語評価極性辞書を用い発言スコアの付与を行う。

発言スコア付与の手順だが、

1. 発言に対し形態素解析エンジン MeCab[6] を用い形態素ごとに分解していく。
2. 分解した形態素それぞれを日本語評価極性辞書と照らし合わせ、スコアを付与する。
3. 形態素ごとの単語スコアを合計し、発言の形態素数で割り平均値を出す。それをその発言のスコアとする。

これらを全発言に対し実行し発言スコアを付与する。

3.2 ネガポジスコアを用いた発言の解析

日本語評価極性辞書で付与したスコアを用い機械学習を行うことで発言から特徴を調べる。本研究では Python の機械学習ライブラリである scikit-learn[7] を使用し、k-近傍法でモデルを生成する。

学習用データには [スコア, ネガティブな単語数, ポジティブな単語数][形態素数] を使用し、正解ラベルを以下にする。

- 性別

  [男性, 女性] の 2 クラス分類。

- 所属

  [民主党, 公明党, 自民党, みんなの党, 共産党, 社民党, 無所属, 維新, 諸派, 所属不明] の 10 クラス分類。

- 市長派

  [市長派, それ以外] の 2 クラス分類。
  市長派とは、地方議会における与野党のことである。地方議会での与党は市長が所属している、または市長を推薦している政党のことであり、その政党に所属している議員をその議会での与党として定義し [市長派] とする。

3.3 生成モデルを使用した議員単位での解析

スコア付与した発言を用いて生成したモデルに議員 1 人の発言それぞれを入力し性別や所属を判別させる。それを全議員に対し実行し、議員ごとの判別結果を学習用データにし教師あり学習を行う。

手順は、

1. ある議員 1 人の全発言を抽出する。
2. 2.2 で生成したモデルにその議員の発言をそれぞれ入力し、[性別] や [所属] を判別させる。
3. これらを全議員に対して実行し、判別結果を出す。

次に判別結果を用い機械学習を行う。このとき、scikit-learn のロジスティック回帰でモデルを生成する。

- 性別

  学習用データに [男性と判別された数, 女性と判別された数, 全体で女性と判別されたものの割合]、正解ラベルに [性別] を設定。[男性][女性] の 2 クラス分類。

- 所属

  学習用データに [各所属と判別された数 (民主党, 公明党, 自民党, みんなの党, 共産党, 社民党, 無所属, 維新, 諸派, 所属不明の 10 ラベル)]、正解ラベルに [所属] を設定。[民主党, 公明党, 自民党, みんなの党, 共産党, 社民党, 無所属, 維新, 諸派, 所属不明] の 10 クラス分類。

4 結果と考察

4.1 日本語評価辞書を用いたネガポジ分析

日本語評価辞書を用いて議員の発言に対してスコアを付与した結果は表 1 となった、

対象	発言数	平均値
全体	256724	0.0094
男性議員	193487	0.0094
女性議員	63237	0.0094

表 1: ネガポジスコアの平均値

平均スコアでは男女で差がないことから、発言での男女差はなかったと考える。

4.2 発言の特徴解析

日本語評価極性辞書でネガポジスコアを付与した発言を使用し発言単位での解析を行った。

- 性別

  [男性 (193487), 女性 (63237)] の 2 クラス、() 内がデータ数である。

  こちらのモデルの精度は 0.74 となったが、混合行列は表 2 となった。

  男女どちらにおいても大半を正しく判別出来ていない。

	男性	女性
男性	171843	21644
女性	45231	18006

表 2: 正解ラベル [性別] での混合行列

- 所属政党

  [民主党 (19926), 公明党 (41240), みんなの党 (2418), 共産党 (60087), 社民党 (1130), 無所属 (68564), 維新 (13029), 自民党 (15523), 諸派 (9360), 所属不明 (25447),] の 10 クラス、() 内が

データ数である。

こちらのモデルでは精度は 0.35 となり、評価指標は表 3 となった。

所属においても評価指標のスコアがどれも低く大半のデータを正しく判別出来ていない。

	スコア
正解率	0.32
再現率	0.34
適合率	0.24
f 値	0.23

表 3: 正解ラベル [所属] での評価指標

- 市長派

  [市長派 (225299), それ以外 (31425)] の 2 クラス、() 内がデータ数である。

  こちらのモデルの精度は 0.87 となったが、混合行列は表 4 となった。

  こちらも大半を正しく判別出来ていない。

	市長派	それ以外
市長派	218837	6462
それ以外	26358	5067

表 4: 正解ラベル [市長派] での混合行列

発言で機械学習を行った場合、どのモデルも大半の発言を正しく判別出来ておらず評価の低いモデルとなった。このことから発言一つ一つには性別や所属などで特徴はないと考える。

4.3 議員の特徴解析

発言を用いて生成したモデルに議員の発言それぞれで性別, 所属を判別させ、その判別結果を用い議員単位での解析を行った。

- 性別

  [男性 (816 人), 女性 (195 人)] の 2 クラス、() 内がデータ数である。こちらのモデルの精度は 0.98 となり混合行列は表 5 となった。

  このモデルでは精度, 評価指標がどちらも高く、ほぼ全てデータを正しく判断出来ている。

	男性	女性
男性	812	4
女性	14	181

表 5: 正解ラベル [性別] での混合行列

- 所属

  [民主党 (70 人), 公明党 (166 人), みんなの党 (7 人), 共産党 (120 人), 社民党 (3 人), 無所属 (224 人), 維新 (71 人), 自民党 (78 人), 諸派 (14 人), 所属不明 (255 人)] の 10 クラス、() 内がデータ数である。

  こちらのモデルの精度は 0.93 となり、評価指標はと表 6 となった。

  こちらのモデルも精度, 評価指標に問題がなく、ほぼ全てを正しく判別できている。

	スコア
正解率	0.93
再現率	0.96
適合率	0.87
f 値	0.90

表 6: 正解ラベル [所属] での評価指標

議員単位で機械学習を行った場合、どのモデルもデータを正しく判別することが出来ており、評価の高いモデルを生成することが出来た。これらから議員それぞれには [性別] や [所属] などの特徴はあると考える。

4.4 まとめ

発言単位で機械学習を用いモデル生成した場合では生成されたモデルの評価は低く、どのモデルも大半の発言を正しく判別することが出来ていなかった。しかし発言単位で生成したモデルを用い、議員単位で機械学習を行った場合に生成されたモデルは精度, 評価指標にも問題がなく、性別, 所属どちらも正しく判別することが出来ていた。

これらのことから、地方議会では一つの発言は性別や所属といった発言議員の属性に応じた特徴を帯びないが、議員それぞれはその属性に応じた特徴のある発言をしていると考えられる。

5　おわりに

本研究では地方議会議事録を自然言語処理と機械学習の手法を用いて、地方議会における発言のネガポジ度合いから性別や所属などで特徴が存在するかどうかを示した。結果として、発言単位では特徴を見つけることは出来なかったが、議員単位では特徴があるということがわかった。

このように本研究では地方議会議事録から特徴抽出を行うための手法として、自然言語処理と機械学習の組み合わせ、解析を行った。更に機械学習から得られた特徴をもとに他の手法を用いて解析を行うことで、様々な解析が可能になると考える。

今後の課題について、所属における [所属不明] の扱いについてがある。[所属不明] とは、本来はその議員の所属政党があったが、なんらかの理由で正しく割り当てることが出来なかったものであり、解析において分類を非常に難しくする要因となるものである。しかしそのような要因がありながらも議員単位での解析においてはうまく分類出来たことから、[所属不明] になっているものの政党を正しく割り当てることができればより精度の高い解析が可能になると考える。

また課題として日本語評価辞書の拡張がある。本研究ではネガポジスコアの付与に日本語評価辞書を用いたが、辞書に載っていない単語が多くあった。そのため辞書の拡張が必要であると考える。拡張を行うことによってスコア付与のできる単語が増えればより精度の高い解析が可能になると考える。

参考文献

[1] 地方議会会議録コーパスプロジェクト – local-politics.jp. http://local-politics.jp/. (Accessed on 01/05/2020).

[2] 名取良太, 岡本哲和, 石橋章市朗, 坂本治也, 山田凱. 地方議会データベースの開発と利用. 情報研究 : 関西大学総合情報学部紀要, Vol. 44, pp. 31–42, aug 2016.

[3] Open resources/japanese sentiment polarity dictionary - 東北大学 乾・鈴木研究室/ communication science lab, tohoku university. https://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2FJapanese%20Sentiment%20Polarity%20Dictionary. (Accessed on 01/07/2020).

[4] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2005.

[5] 東山昌彦, 乾健太郎, 松本裕治. 述語の選択選好性に着目した名詞評価極性の獲得. 言語処理学会第 14 回年次大会論文集, pp. 584–587, 2008.

[6] Mecab: Yet another part-of-speech and morphological analyzer. http://taku910.github.io/mecab/. (Accessed on 01/05/2020).

[7] scikit-learn: machine learning in python — scikit-learn 0.22.1 documentation. https://scikit-learn.org/stable/. (Accessed on 01/05/2020).