都道府県議会会議録からの意見や意志を表す発言の抽出

坂地 泰紀[1] 酒井 浩之[1] 小林 暁雄[2] 内田 ゆず[3] 乙武 北斗[4] 高丸 圭一[5] 木村 泰知[6]
成蹊大学[1] 豊橋技術科学大学[2] 北海学園大学[3] 福岡大学[4] 宇都宮共和大学[5] 小樽商科大学[6]
{hiroki_sakaji, h-sakai}@st.seikei.ac.jp, a-kobayashi@cs.tut.ac.jp, yuzu@eli.hokkai-s-u.ac.jp,
ototake@fukuoka-u.ac.jp, takamaru@kyowa-u.ac.jp, kimura@res.otaru-uc.ac.jp

1 はじめに

地方議会会議録には，地方議会における発言が書き起こされ，記録されている．地方分権や地方創成などの議論が盛んになる中，地方議会会議録の分析が重要性を増している．地方政治学等の分野において，会議録の調査・分析が行われはじめている [1].

会議録には，意見や意志を表した発言が含まれている．例えば,「飼料用米など非主食用米の作付推進や収量の向上が重要な課題となると考えております。」という発言は主に意見を表している．このような発言を抽出することができれば，誰がどのような意見を持っているかを有権者が知ることができ，選挙時の候補者選びに役立つ．また，過去の会議録から上記のような発言が抽出できれば，実際にその意見に対して何かしらの対策が行われたのかという調査も可能となる．

現在，全ての都道府県議会が会議録を Web 上に公開している．しかしながら，それぞれの自治体が個別に公開しており，フォーマットも統一されていないため，全国規模の調査･分析を行うには，収集作業やデータ整理に労力がかかるという問題がある．そこで我々は，地方議会会議録を用いた全国規模の研究を推進することを目指して，**地方政治コーパス**の構築を進めている [2]．本研究では，収集した地方政治コーパスを用いて，意見や意志を表す発言を自動的に抽出する手法を開発する．

2 意見・意志表明発言

本研究で扱う意見・意志を表す発言は，意見や今後取り組む予定である事案について発言したものである．本研究では，意見・意志を表す発言を**意見・意志表明発言**と定義する．意見・意志表明発言の例を表 1 に示す．表 1 において，主に一つ目の発言が意見を表し，2 つ目の発言が意志を表している．

表 1: 意見・意志表明発言の例

- 災害が多発する本県でも、施設整備を中心としたハード対策、情報提供などのソフト対策の両面から防災対策を強化する必要があると考えますが、即効的な取組として情報提供などソフト対策の充実が必要ではないかと考えます。
- 今後、適宜庁内関係部局で構成いたします会議も開催をいたしまして、当面の対応策についての全体的な取りまとめを行うなど、新たなエネルギー社会づくりに向けた取り組みを進めてまいります。

3 意見・意志表明発言の抽出手法

本章では，意見・意志表明発言抽出手法について述べる．本研究では，意見・意志表明発言を抽出するための手がかりとなる表現 (**手がかり表現**) を用いて，意見・意志表明発言を抽出する．例えば,「と考えます。」などが手がかり表現となる．しかしながら，手がかり表現の数は多く，全てを人手で網羅するのは難しい．そこで，ブートストラップ手法 [3] を用いて，手がかり表現の獲得を行う．ブートストラップ手法では，手がかり表現を獲得する際に，手がかり表現の直前に良く出現する表現 (**共通頻出表現**) を用いる．本手法を Algorithm 1 に示す．

3.1 共通頻出表現の獲得

本節は，Algorithm 1 での *get_commons* にあたる．本手法では，適切な共通頻出頻出表現を獲得するために，まず共通頻出表現の候補 (共通頻出表現候補) を獲得する．図 1 に示すとおり，係り受け解析器を用いて文を係り受け解析する．その後，手がかり表現を構

Algorithm 1 Bootstrapping

Input: Initial Clue Set $C = (c_0, c_1, ..., c_n)$ and Iteration Number T

Output: Opinion Sentences O

1: $\acute{C} \leftarrow C$, $M \leftarrow \emptyset$ ▷ M is Common Set
2: **for each** $t \leftarrow 1...T$ **do**
3: $M \leftarrow M + get\_commons\left(\acute{C}\right)$
4: $\acute{C} \leftarrow \acute{C} + get\_clues\,(M)$
5: **end for each**
6: $O \leftarrow get\_opinions\left(\acute{C}\right)$
7: **return** O

成する最初の文節から，手がかり表現を構成する文字を除いた文字列を共通頻出表現候補として獲得する．図 1 では，手がかり表現を構成する最初の文節が「まいりたいと」となり，この文節から手がかり表現を構成する文字である「と」を除去し,「まいりたい」を共通頻出表現候補として獲得する．

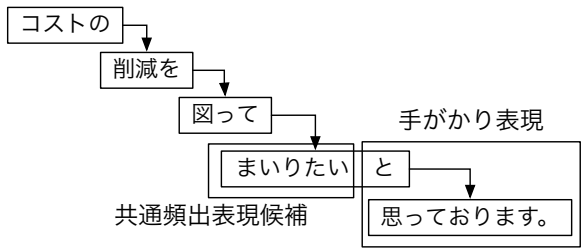

図 1: 共通頻出表現候補獲得の例

共通頻出表現候補の中には，共通頻出表現として適切でないものも含まれている．そこで，様々な手がかり表現とまんべんなく共起する共通頻出表現候補は適切であるという仮定に基づき，手がかり表現との共起確率に基づくエントロピーを用いて選別を行う．エントロピーは，事象が等確率で発生するとき，最も高い値をとる．逆に，事象によって発生確率が偏っている場合は，低い値となる．この特性を利用し，様々な手がかり表現とまんべんなく共起する場合に高い値となり，共通頻出表現として獲得される．また，本手法はブートストラップ手法であるため，エントロピーの値が低い共通頻出表現候補は，特定の手がかり表現としか共起しないものであることから，これを獲得してしまうと，local optimum に落ちてしまう可能性がある．2 回以上出現した共通頻出表現候補に対して，以下の式 1 を用いてスコアを計算する．なお，スコアは 0 から 1 の値を取るように正規化している．

$$Score(m) = \frac{H(m)}{max_m H(m)}, \tag{1}$$

$$H(m) = -\sum_{c \in C(m)} P(c,m) \log_2 P(c,m), \tag{2}$$

$$P(c,m) = \frac{f(c,m)}{\sum_{c' \in C(m)} f(c',m)}, \tag{3}$$

ただし，

$C(m)$: 共通頻出表現候補 m と共起する手がかり表現の集合

$max_m H(m)$: すべてのエントロピー $H(m)$ の中で最大のもの

$P(c,m)$: 手がかり表現 c と共通頻出表現候補 m が共起する確率

$f(c,m)$: 手がかり表現 c と共通頻出表現候補 m の共起数

$Score(m)$ が閾値 α 以上の共通頻出表現候補を共通頻出表現として獲得する．以下に，獲得した共通頻出表現の例を示す．

重要なのではないか　注目していきたい
対応していくべきだ　手段ではないか　原則だ

3.2 手がかり表現の獲得

本節は，Algorithm 1 での *get_clues* にあたる．図 2 で示すように，共通頻出表現の直後の格助詞から文末までを手がかり表現候補として獲得する．図 2 において,「.」はワイルドカード,「*」は 0 回以上の繰り返しを意味する．

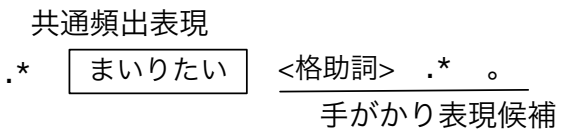

図 2: 手がかり表現候補獲得の例

また，手がかり表現候補の中にも不適切なものが含まれるため，選別を行う．2 回以上出現し，かつ，3 文字以上の手がかり表現候補に対して，共通頻出表現と同様にエントロピーを用いて手がかり表現を獲得する．以下に，獲得できた手がかり表現の例を示す．

と思うわけでございます。　と感じています。
と私は思っています。　と思うわけです。　と思う。

3.3 意見・意志表明発言の抽出

本節は，Algorithm 1 での *get_opinions* にあたる．反復処理により獲得した手がかり表現集合を用いて，手がかり表現を含む文を意見・意志表明発言として抽出する．ただし，本手法では手がかり表現集合に「まいります。」を追加した状態で意見・意志表明発言を抽出する．「まいります。」は意見・意志表明発言を獲得するうえで重要な手がかり表現となるが，本手法で用いたブートストラップではうまく処理することができなかったため，本処理において追加することとした．

4 評価実験

学習データとしてランダムに抽出した 600 発言を用い，学習データ以外からランダムに抽出した 400 発言を評価データとして用いる．人手で評価したところ，学習データ中には 100 の，評価データには 78 の意見・意志表明発言が含まれていた．本手法では，学習データと評価データ以外を手がかり表現の獲得に用い，閾値と反復回数を決定するために学習データを用いる．初期手がかり表現には，「と思います。」「と思っております。」「と考えております。」「と考えます。」の 4 個を用いた．

比較手法として，SVM と Random Forest(RF)，Neural Network(NN) を用いる．SVM は scikit-learn[1] を用い，カーネルは線形，C の値は 1 とする．SVM の素性には，形態素ユニグラムと形態素バイグラムを持ちいるものと，word2vec[2] を用いるものの 2 種類を実験する．RF も SVM と同様に scikit-learn を用い，木の数は 50 とする．RF の素性には，形態素ユニグラムと形態素バイグラムを用いる．NN は keras[3] で実装し，素性には同じく形態素ユニグラムと形態素バイグラムを用いる．形態素解析器としては Mecab[4] を用い，係り受け解析器には CaboCha[4] を用いた．

5 評価結果

本手法と比較手法を評価した結果を表 2 に示す．表 2 において，本手法は閾値 0.8，反復回数 6 の場合を用いている．図 3 に本手法を学習データに対して閾値 0.8 で反復させた場合の評価結果を示す．図 3 において，baseline は初期手がかり表現だけを用いて意見・意志表明発言を抽出した場合の結果である．

本手法を学習データに対して適用した場合，各閾値において最も高い F 値を得た反復回数時の結果を表 4 に示す．表 4 において，各横軸には，閾値と反復回数を示している．また，本手法によって抽出できた意見・意志表明発言の例を表 3 に示す．

表 2: 評価結果

	適合率	再現率	F 値
本手法	0.79	0.57	0.66
SVM(uni-bigram)	0.69	0.54	0.60
SVM(word2vec)	0.46	0.41	0.43
RF	0.83	0.44	0.57
NN	0.64	0.47	0.53

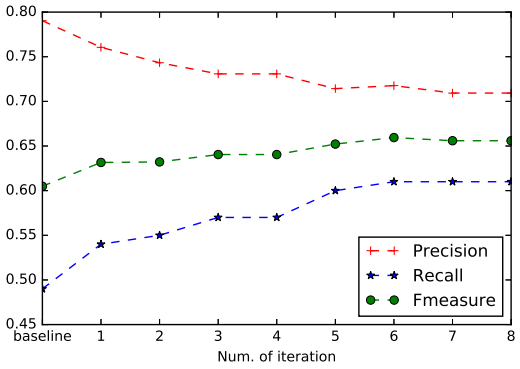

図 3: 本手法を学習データに対して閾値 0.8 で反復させた場合の結果

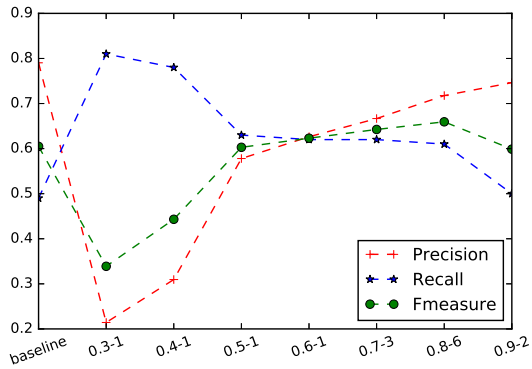

図 4: 閾値を変更しながら手がかり表現を獲得し，学習データに対して適用した結果

6 考察

表 2 より，本手法の結果が他の手法に比べ最も高い F 値となった．適合率に関しては，RF が最も高い値

[1] http://scikit-learn.org/stable/
[2] https://radimrehurek.com/gensim/
[3] https://keras.io/
[4] http://code.google.com/p/mecab/

表 3: 抽出した意見・意志表明発言の例

目標にはなかなか届かない状況ではありますけれども、引き続きこのような取り組みを続けまして、ジェネリック医薬品についての正しい知識の普及啓発に努めてまいりたいと思います。
今後とも台風２号の影響による農家支援については、農家の要望等を踏まえ、関係機関等と連携し対応を検討したいと考えております。
道といたしましては、道教委と連携し、小中学生が外国の文化に触れる機会や、高校における外国語学習などを通じて、児童生徒の国際理解やコミュニケーション能力の向上が図られるよう努めてまいります。

を達成したものの，再現率が 0.44 と低かった．本手法は，適合率では RF に劣るものの，再現率が 0.57 と最も高く，その結果，全手法の中で最も高い F 値を達成した．以上のことから，本手法の有効性を確認することができた．

図 3 より，反復回数が増えるほど，適合率が下がり，再現率が向上している．また，図 4 より，閾値が下がるほど，適合率が下がり，再現率が向上している．このような適合率と再現率の推移は，ブートストラップ手法全般に見られる．本手法もブートストラップ手法であるため，上記のような適合率と再現率の推移となったと考える．

表 2 と図 3，図 4 より，baseline の F 値が比較手法で用いた他の機械学習手法より高いことが示されている．この結果から，本タスクのおいては，機械学習手法よりもキーワードによるパターンマッチングの方が適していると考えることができる．また，baseline と本手法との F 値の差が大きくないため，本手法の更なる改善が求められる．

7 関連研究

ブートストラップを用いた研究としては酒井ら [5] や Pantel et al.[6] の研究がある．酒井らは，企業の決算短信 PDF から業績要因を表す文を自動的に抽出する手法を提案している．Pantel et al. の提案している手法は，パタン p を用いて 2 つの名詞対 $i = x, y$ を獲得する手法である．それに対して，我々が提案する手法は文末に着目したブートストラップ手法となっている点が異なる．地方議会会議録を対象とした研究としては，木村ら [7] の研究がある．木村らは，主辞に着目した地方政治問題の抽出手法を提案している．それに対して，本研究では地方政治問題ではなく，意見・意志表明発言を抽出する研究となっている．

8 おわりに

本研究では，都道府県議会議事録から意見・意志表明発言を抽出する手法を提案した．本手法はブートストラップ手法となっており，エントロピーを用いて不適切な表現をフィルタリングした．評価の結果，比較手法よりも本手法の F 値が高くなり，本手法の有用性を示すことができた．今後は，意見・意志表明発言と発言者の紐づけを行い，発言検索できるようにしたい．

謝辞

本研究は，JSPS 科研費 16H02912 と 15K00315 の助成を受けたものです．

参考文献

[1] 増田正. 地方議会の会議録に関するテキストマイニング分析 : 高崎市議会を事例として. 地域政策研究, Vol. 15, No. 1, pp. 17–31, aug 2012.

[2] 田中琢真, 小林暁雄, 坂地泰紀, 内田ゆず, 乙武北斗, 高丸圭一, 木村泰知. 地方政治コーパス構築に向けた都道府県議会会議録からの発言データの抽出. 第 32 回ファジィシステムシンポジウム (FSS2016), pp. 251–254, 2016.

[3] 坂地泰紀, 酒井浩之, 増山繁. 決算短信 pdf からの原因・結果表現の抽出. 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822, 2015.

[4] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[5] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀. 企業の決算短信 pdf からの業績要因の抽出. 人工知能学会論文誌, Vol. 39, No. 1, pp. 172–183, 2015.

[6] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *in Proceedings of the 21st Internatinal Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 113–120, 2006.

[7] 木村泰知, 関根聡. 主辞に基づく政治問題抽出手法. 人工知能学会論文誌, Vol. 28, No. 4, pp. 370–378, 2013.