Gaussian LDA を用いた地方議会会議録のトピック分析

佐々木 稔
茨城大学工学部情報工学科
minoru.sasaki.01@vc.ibaraki.ac.jp

木村 泰知
小樽商科大学商学部社会情報学科
kimura@res.otaru-uc.ac.jp

1 はじめに

本論文では，地方議会において，個別の議員がどのような内容についての発言を行ったかについての分析を行う．

現在，数多くの自治体で情報公開への取り組みの一つとして，議会での発言を書き起こした会議録が公開されている．このような地方議会会議録を収集し，コーパスとしての整備も進められたことで，学際的な利用も可能となっている [4][5]．そのため，情報抽出や議会活動の可視化など，このコーパスを利用した様々な研究が行われている．

本研究では，会議録コーパスに対して，テキストマイニング手法を利用して地方議会でどのような話題が議論されているかについて分析を行う．地方議会で議論される話題の分析はこれまでに単語の頻度や共起などの情報を手がかりとした手法が存在する [6][7][8]．本研究においても，地方議会において何が話題となっているのかを明らかにすることが目標である．例えば，図 1 において，[発言内容] のようなオリンピック・パラリンピック招致に関連する発言があった場合，[話題の代表語] にある単語の集合を求めることが課題となる．従来研究では，単語，複合名詞の出現頻度や共起頻度を比較して分析を行う方法と Ward 法を用いて階層的に単語のクラスタリングを行って分析する方法が存在する．これらの方法に対して，考慮すべき問題が 2 点あると考えられる．その 2 つの問題をまとめると以下のようになる．

- 分析を行う際に，最も適切な単語単位はどの程度なのか
- 他の都道府県議会会議録から得られた単語の分散表現は利用可能なのか

本稿では，これら 2 つの問題に対処するため，地方議会の話題分析における最適な単語単位と分散表現の利用可能性について調査を行う．話題分析に使用する単語単位の種類として，MeCab で利用可能な辞書である ipadic, unidic, mecab-ipadic-NEologd および Comainu を利用する．得られた単語集合に対して，単語の分散表現を用いたトピック分析手法である Gaussian LDA (Latent Dirichlet Allocation) を用いて，出現単語からトピックを抽出する．使用した各辞書に対して，Gaussian LDA により得られたトピック集合を比較することで，話題の分析においてどの辞書を用いた場合が単語単位として最も適切なのかを検証する．また，単語単位を unidic に限定し，他の都道府県議会会議録から得られた分散表現と国語研日本語ウェブコーパスから得られた nwjc2vec の分散表現を用いて得られたトピック集合を比較して，他の都道府県議会会議録から得られた単語の分散表現の有効性を検証する．

[発言内容]
委員会では、継続的にオリンピック・パラリンピック招致のスケジュールを確認するとともに、招致活動の現状について報告を聴取し、質疑を行っております。

[話題の代表語]
オリンピック パラリンピック 招致 招致活動

図 1: 話題 (オリンピック招致) における代表的な単語

2 会議録からのトピック分析方法

2.1 使用データと発言の抽出

本研究では，東京都議会を対象としてトピック分析を行う．東京都議会のウェブサイト[1]で公開されている平成 23 年 4 月から平成 27 年 3 月までの本会議会議録のテキストデータを利用する [3]．このデータは会

[1] https://www.gikai.metro.tokyo.jp/record/proceedings/

議録の内容を句点で区切り，1 文単位に分割をして保存されている．しかし，このデータには議員の発言だけではなく，その場の状況説明 (「(拍手)」や「〔知事 (人名) 登壇〕」など)，添え書き (「(パネルを示す)」など) や朗読動作を表す「非発言」部分が存在する．会議録データの「発言」部分と「非発言」部分を切り分けて「発言」部分のみを抽出することで，全体として 37,651 文の発言を使用した．この発言に対して，定例会における各議員が発言した文集合を文書単位とすることで，327 件の発言文書を作成した．

2.2 発言データの単語分割

発言文書のすべての文に対して，異なる 4 種類の辞書を用いて形態素解析を行い，単語に分割する．形態素解析器は MeCab を利用し，辞書には以下の 4 種類を使用する．

- ipadic
  IPA 品詞体系に基づき構築された標準的な辞書で，MeCab の辞書として一般的に使われる．
- unidic[2]
  国立国語研究所が開発した，斉一な言語単位である短単位で書き言葉文書を自動解析するための辞書である．
- mecab-ipadic-NEologd (neologd)[3]
  IPA 辞書を拡張した新語や固有表現を効果的に抽出できる辞書である．
- Comainu[4]
  複数の短単位単語を結合した長単位解析を行うための辞書である．イベント名や委員会名など複数の名詞連続を効果的に抽出することができる．

上記の 4 種類の辞書を用いて，形態素解析を実行した結果から名詞語句のみを取り出す．抽出した名詞語句を単語列として発言文書を表現し，この単語列を用いてトピック分析を行う．

2.3 Gaussian LDA によるトピック抽出

名詞単語列として表した発言文書に対して、Gaussian LDA を利用してトピックを抽出する。Gaussian LDA は入力された文書中に存在する潜在的なトピックを自動で抽出するモデルで、単語の分散表現を用いることで単語間の関連性を考慮できるモデルである。

この Gaussian LDA を用いて，単語列として表した発言文書からトピックの抽出を行う．Gaussian LDA で得られるトピックは高頻度の語句に対してはトピックにまとまりにくい傾向があるため，全発言文の 75% 以下の割合で出現する語句を対象としてトピック抽出を行った．Gaussian LDA はトピック数と学習回数をあらかじめ設定する必要があり，本稿ではトピック数を 50，トピックの学習回数を 100 回と設定し，トピック抽出を行った．

2.4 単語の分散表現データ

Gaussian LDA を利用してトピック分析を行う際，出現単語をベクトルで表現した分散表現が必要となる．単語の分散表現は別に用意した大規模な文書集合から，word2vec などを用いて求めたものを用意する．本稿では他の都道府県議会会議録から得られた分散表現と国語研日本語ウェブコーパスから得られた nwjc2vec の分散表現 [1] の二種類を用いる．

他の地方議会会議録の分散表現は，[5] において得られた地方議会会議録コーパスに対して word2vec を用いて得られる各単語のベクトル表現である．使用データの単語分割と同じ 4 種類の辞書を用いて形態素解析を行い，発言を単語分割した単語列を抽出する．この単語列を求める場合は単語の品詞を限定せず，すべての出現単語を空白で区切った単語列を求める．会議録コーパスの単語列に対して，word2vec の CBOW モデルを用いて各単語を 200 次元のベクトルで表現した分散表現を求める．このとき，word2vec のパラメータは最小出現頻度を 20，ウィンドウ幅を 201 と設定し，残りのパラメータはデフォルト値を使用する．

nwjc2vec は国立国語研究所が公開する分散表現のデータである [1]．nwjc2vec はウェブを母集団とした約 258 億語からなる日本語コーパスで，nwjc2vec はこのコーパスに対して word2vec の CBOW モデルを用いて各単語を 200 次元のベクトルで表現されている．nwjc2vec を作成した際のパラメータはウインドウ幅が 8 となっており，地方議会会議録から得られた分散表現とは値が異なっている．nwjc の単語分割には unidic が用いられているため，実験では辞書を unidic に合わせた場合にどちらの分散表現が有効かどうかを検証する．

[2] https://unidic.ninjal.ac.jp/
[3] https://github.com/neologd/mecab-ipadic-neologd
[4] http://comainu.org/

3　トピック分析

3.1　辞書の違いによるトピックの比較

ipadic, unidic, neologd, Comainu の各辞書で発言文書を単語分割した 4 種類のデータから得られたトピックを比較する．

ipadic を用いた場合，表 1 に示すように，多くの議員が話題にした東日本大震災に関連する単語については，一部関連しない単語が含まれるものの，トピックとしてまとまっていると考えられる．しかし，その他のトピックについては，表 3 にある ipadic のトピックのように，一つの内容を表現したトピックであるとは考えにくく，様々な話題の単語を集めたトピックが得られる結果となった．複数の話題がトピックとして集まる傾向は unidic や neologd を用いた場合でも同様であった．この要因として，これらの辞書を用いると単語が短く分割されるため，複数の内容で使われる一般的な単語として出現しやすくなることが挙げられる．例えば，表 1 のトピック 36 にある「浸水」は，発言データにおいて津波による浸水被害，豪雨による浸水被害，および，浸水対策という複数の話題で使われている．しかし，トピック 36 では「震災」,「基盤」や「リスク」が存在することから，すべての内容が含まれたトピックになっていると考えられる．

Comainu を用いた場合に得られたトピックの一部を表 2 に示す．長単位の単語でトピックが構成されているため，人が確認しやすい内容となっている．トピック 3 は，ある議員が一度に行った発言の中に「都心と臨海副都心とを結ぶ公共交通について」,「認定こども園について」,「予算編成と財政運営について」をすべて含んでいたことから，ひとつのトピックとしてまとめられたと考えられる．トピック 5 でも，1 人の議員の発言に「災害時の病院」,「高齢化と地方分権」,「豊洲の土壌汚染問題」があり，それらがトピックとなっている．このように，数人が共通の話題について発言を行った内容がトピックとしてまとまる傾向がある．複数の話題が含まれているが，トピック分析を行うには Comainu の単語分割が最も有効ではないかと考えられる．しかし，Comainu を用いる場合，他の辞書を用いる場合と比較して未知語が数多く存在する．Comainu で得られた長単位の名詞語句のうち，地方議会会議録の分散表現に存在しない語句の数は 16,593 語で，ipadic を用いた場合の 1,549 語よりも多かった．この中には「東京オリンピック・パラリンピック」といった総称や「東京スカイツリー」などの東京独自の語句が存在する．低頻度ではあるものの，これらの語句を有効活用するための改良は今後の課題とする．

表 1: ipadic を用いた場合のトピックの一部

トピック 25(ipadic)		トピック 36(ipadic)	
単語	頻度	単語	頻度
被災	391	確保	783
特別	377	震災	323
条例	347	地方	272
措置	148	解決	184
認定	117	適切	171
福島	99	基盤	167
軽減	87	津波	165
固定	56	浸水	120
放射能	54	視察	116
法律	46	誘致	115
放射線	43	確実	104
徴収	40	若年	77
領土	39	リスク	74
生物	38	的確	68

表 2: Comainu を用いた場合のトピックの一部

トピック 3	トピック 5	トピック 20	トピック 41
交通	魅力	対策	視野
両面	分野	子供	趣旨
協定	病院	子供達	災害対策
財政運営	割合	連携	水害
公共交通	工事	被災地	暫定措置
保育園	最初	防災対策	水門
児童虐待	行動	開催	超高齢社会
バス	精神	行政	幼児教育
商業施設	地方	減少	形成
説明責任	イベント	河川	気候変動
再開発	条例案	国内外	高齢者施策
議事	地方分権	従来	教室
財政	日程	人材育成	在宅医療
最小限	橋梁	意思	医療機能
執行	少子化	先頭	ガイドライン
待機児童対策	人員	コスト	数々
決算審査	人口減少	議員	高齢化社会

3.2　会議録分散表現と nwjc2vec の比較

都道府県議会会議録から得られた分散表現と nwjc2vec のそれぞれについて，Gaussian LDA で得られたトピック集合を比較し，トピックの検証を行う．議論を代表する単語を予め設定し，その単語を含むトピックを比較して検証を行う．

例として，豊洲市場移転問題で使われる「移転」を検証単語として設定する．この「移転」を含むトピックを表 4 に示す．会議録の分散表現を用いた場合,「移転」と同じトピックに汚染，商店，液状や水質といった豊洲移転問題で出現する単語を含んでいる．nwjc2vec を用いた場合，影響，協議や更新といった前後に出現しやすい単語がトピックとしてまとまり，豊洲移転問題とは直接関係のない単語を多く含んでいる．この結果から，会議録の分散表現を利用する方が有効なトピッ

表 3: 単語「下水道 (下水)」を含むトピック

トピック 48(ipadic)		トピック 47(unidic)		トピック 32(neologd)		トピック 14(Comainu)	
単語	頻度	単語	頻度	単語	頻度	単語	頻度
対策	1768	日本	1420	交通	308	東京都	552
重要	1325	所見	764	**下水道**	186	耐震化	171
課題	768	活用	718	子供	182	九月	97
大震災	454	期待	340	患者	170	建物	80
規模	322	子供	314	うち	126	サービス	77
下水道	241	子ども	304	具体的	98	**下水道**	55
認知	215	経営	300	需要	87	利便性	44
都政	201	**下水**	295	段階	77	助成	36
都議会	168	一方	251	実効性	63	意向	31
明らか	143	人口	155	本格	62	管理	29
行動	108	労働	140	危険	54	私立学校	28
台風	76	留学	62	悪化	41	競技	26
寄与	69	アメリカ	52	登録	37	下水道事業	23
徹底	61	規定	48	地下	32	下水道施設	22

クを抽出することが可能であると考えられる．

移転問題に関する単語が同じトピックに集まるのは，使用する分散表現の違いと分散表現作成時のウインドウ幅の違いが要因として挙げられる．会議録分散表現では「移転」の類似単語として「建物」が上位に存在するが，nwjc2vec では類似度が低かった．都議会の会議録には「建物」の液状化現象や汚染に関する記述があるため,「移転」の関連語として分散表現が有効に働いていると考えられる．また，単語列全体を範囲とするようにウインドウ幅を設定することで，同じ議論内容において共起する単語の組み合わせを捉えることができるため，関連する単語が同じトピックに集まりやすい傾向がある．ウインドウ幅が狭い場合は「移転協議」や「移転の影響」などのように，近い場所で共起する名詞単語が同じトピックに集まる傾向がある．ウインドウ幅の広さを大きくすることで，議論の内容をトピックとしてまとまりやすくなると考えられる．

4 おわりに

本稿では，東京都議会を対象としてどのような話題が議論されているかについて，Gaussian LDA を利用してトピック分析を行った．ipadic, unidic, mecab-ipadic-NEologd, Comainu の各辞書で発言文書を単語分割したデータに対してトピックの比較を行った結果，ipadic, unidic, mecab-ipadic-NEologd は内容ごとにまとまったトピックを得ることが難しかったが，Comainu では他の辞書と比べてまとまったトピックが得られる結果となった．他の地方議会会議録から得られた分散表現を用いることも，東京都議会のトピックを分析する際に効果があることを確認することができた．しかし，現状において本手法は話題のまとまり

表 4: 単語「移転」を含むトピック

トピック 31(会議録)		トピック 15(nwjc2vec)	
単語	頻度	単語	頻度
全体	296	影響	309
困難	270	協議	144
世代	193	**移転**	123
職員	181	更新	100
汚染	127	まま	89
移転	123	関心	84
国家	121	感染	83
商店	97	研修	46
液状	88	スタート	44
学習	85	波及	41
候補	66	職場	40

をうまく捉えていないことが多い．そのため，話題に関連する語句がトピック内に多く存在し，議論の内容をまとめられるように改良する予定である．

参考文献

[1] Masayuki Asahara. NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, Vol. 24, No. 2, pp. 7–25, Feb. 2018.

[2] Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian lda for topic models with word embeddings. In *ACL (1)*, pp. 795–804. The Association for Computer Linguistics, 2015.

[3] 桧森拓真, 木村泰知, 坂地泰紀. 地方議会会議録における発言文の推定. 言語処理学会第 24 回年次大会, pp. 125–128, 2018.

[4] 齋藤誠, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則. 地方議会会議録の収集とコーパスの構築. 言語処理学会第 17 回年次大会, pp. 368–371, 2011.

[5] 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, pp. 251–254, 2016.

[6] 内田ゆず, 高丸圭一, 乙武北斗, 木村泰知. 都道府県議会会議録コーパスを用いた議員の議会活動の可視化に向けて. 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 1E3–03, 2018.

[7] 高丸圭一. 地方議会では何が話題になっているのか―宇都宮市議会会議録のテキストマイニング―. 都市経済研究年報, Vol. 13, pp. 162–173, 2013.

[8] 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.