

大規模地方議会会議録の分散表現を用いた地方議会のトピック分析

Topic Analysis in Local Assemblies using Word Embeddings Obtained
from Large-scale Local Assembly Minutes

佐々木稔 ^{*1}

Minoru Sasaki

乙武北斗 ^{*2}

Hokuto Ototake

木村泰知 ^{*3}

Yasutomo Kimura

茨城大学 ^{*1}

Ibaraki University

福岡大学 ^{*2}

Fukuoka University

小樽商科大学 ^{*3}

Otaru University of Commerce

In our present study, we analyze what topics are discussed in local assembly using text mining methods. Although there have been several studies to analyze discussion using topic model, existing studies does not evaluate topics for words obtained using several word segmentation dictionaries and have not discussed about the effectiveness of word embeddings obtained from large-scale training data. In this paper, we obtain the topics for the NTCIR14 Segmentation task data set by using Embedded Topic Model(ETM) that either uses the word embeddings trained on Regional Assembly Minutes Corpus or Wikipedia articles. Then, we compare the topics obtained by the two topic models. Experimental results show that, by using the Comainu dictionary, we can easily understand the meaning of the topic and easily assign labels to topics. However, we can not confirm a clear difference between the topic models trained on Regional Assembly Minutes Corpus and Wikipedia articles.

1. はじめに

近年、テキストマイニングの分析ツールとして、トピック分析が利用されている[4]。トピック分析は、文書内で共起する単語の顕在的共起性や文章上に現れない潜在的共起性を考慮しながら、文書に含まれる各単語にトピックを割り当てることが一般的である。代表的な確率的トピックモデルである Latent Dirichlet Allocation(LDA)は、ディリクレ分布を用いて文書集合の中にどのような単語集合からなるトピックがあるかを出力する。他にも、トピックモデルは、潜在的意味を推定することを目指して、Embedded Topic Modeling (ETM) やニューラルトピックモデルなどの様々な手法が提案されている。

しかしながら、トピック分析の出力となる単語集合の分布は、単語単位（つまり、単語分割の問題）や学習データの違いにより、結果が異なり、トピック分析として使用しづらい場合がある。また、日本語の単語分割については、対象分野によって適切な分割方法が異なることから、解決する問題を想定して選択する必要がある。既存研究では話題の分析を行う際に、最も適切な単語単位はどの程度なのか、大規模な都道府県議会会議録から得られた単語の分散表現が利用可能なのかについて研究が行われていない。

そこで、本研究では、地方議会における議員の発言を対象として、質問に含まれるトピックを同定することを最終目標として、トピック分析の課題解決を試みる。具体的には下記の2つの問題がある。

1. 単語分割の問題：単語分割は、形態素解析ツールと辞書の組み合わせによって異なる。その単語分割の違いは、単語集合の分布の違いにもつながり、有効なトピック分析ができるのかにも影響する。
2. 分散表現の学習コーパスの問題：分散表現を学習する場合、Wikipedia以外に利用できる学習コーパスが少ないという問題がある。本研究では、大規模な地方議会会議録を学習データがどの程度利用可能なのかを明らかにする。

連絡先: 佐々木稔、茨城大学、〒316-8511 茨城県日立市中成沢町4-12-1、minoru.sasaki.01@vc.ibaraki.ac.jp

本稿では、地方議会会議録に対して、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テキストマイニング手法を利用した分析を行う。具体的には、NTCIR14 Segmentation taskで利用されたデータセット^{*1}を用いて、単語分割、および、学習データの違いにより、トピックモデルの結果がどの程度異なるのかを明らかにする。単語分割については長単位解析ツールComainuを利用する。学習データについては、Wikipedia、および、地方議会会議録を学習データとした分散表現をもいた結果を比較することで違いを明らかにする。さらに、大規模地方議会会議録から学習した分散表現は <http://local-politics.jp>において公開する。

2. 分散表現

本研究では、分散表現を作成する単語の単位として、長単位解析ツールComainu^[5]を用いる。Comainuは複数の短単位単語を結合した長単位解析を行うためのツールであり、地方議会会議録に多く含まれると思われ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。

分散表現のモデルは、Mikolovら[2]のskip-gramモデルを用いて構築した。実装はPythonのgensimライブラリ^{*2}に含まれるmodels.word2vecモジュールを用い、表1で示すパラメータを与えてモデルの構築を行った。

モデルの構築材料となるコーパスは、Wikipedia日本語版の記事、およびWeb上に公開されている地方議会会議録の2種類を用い、それぞれのコーパスから独立したモデルを構築した。以下、コーパスの詳細、および構築したモデルの比較について述べる。

2.1 Wikipedia

2020年1月25日における最新のWikipedia日本語版ダンプデータ^{*3}を利用した。総単語数は約9億6000万語、異なる単語数は約3500万語である。構築した分散表現モデルの語彙数は3,399,560となった。

*1 <https://github.com/kmr-y/NTCIR14-QALab-PoliInfo-FormalRunDataset>

*2 <https://radimrehurek.com/gensim/index.html>

*3 <https://dumps.wikimedia.org/jawiki/latest/>

表 1: word2vec のモデル構築パラメータ

パラメータ	値
size	200
window	5
negative	5
min_count	5
sg	1 (skip-gram)
hs	0
iter	20

表 3: 「財源」との類似度上位 10 件の単語

モデル	単語
Wikipedia	税収, 国庫, 歳入, 支出, 原資, 税金, 補助金, 資金, 運営費, 交付金
地方議会会議録	一般財源, 特定財源, 必要財源, 自己財源, 歳入財源, 財源確保, 補正財源, 独自財源, これら財源, 所要財源

表 2: 「高齢者」との類似度上位 10 件の単語

モデル	単語
Wikipedia	障害者, 乳幼児, 身体障害者, 介護者, 75 歳以上, 交通弱者, 無職世帯, がん患者, 乳幼児連れ, お年寄り
地方議会会議録	お年寄り, 高齢者等, 高年者, 老人, ひとり暮らし, 一人暮らし, 障害者, 独居老人, おとしより, ひとり暮らし高齢者

2.2 地方議会会議録

会議録をウェブ公開している自治体は、システム開発業者に委託していることが多い、主に 4 社（会議録研究所、大和速記情報センター、フューチャーイン、神戸総合速記）の会議録検索システムが採用されている。菅原らは、618 の自治体が上記の 4 社の会議録検索システムを利用していることを報告している [3]。

本研究では、主に上記 4 社の会議録検索システムを用いて公開されている会議録を対象に、プログラムによる自動処理にて収集、データベース化された地方議会会議録コーパス [6][7] を利用した。今回の分散表現作成に用いたデータは、コーパスに登録されている、すべての都道府県を網羅した 425 自治体（都道府県 20、市町村 392、特別区 13）の会議録である。本データは自治体によって偏りがあるが、1947 年から 2011 年までの発言が含まれており、発言数は約 1 億 3000 万、総単語数は約 38 億語、および異なり単語数は約 1800 万語である。構築した分散表現モデルの語彙数は 3,116,959 となった。

2.3 比較

構築した分散表現モデルのコーパスによる違いを比較するために、「高齢者」の分散表現とコサイン類似度が高い単語上位 10 件を、それぞれのモデルで出力した結果を表 2 に示す。 Wikipedia モデルでは「高齢者」と類似している単語として、「障害者」「乳幼児」「介護者」など、類似した文脈で出現する別の意味の単語が含まれている。一方で地方議会会議録モデルの場合、「お年寄り」「老人」のような類似した意味の単語が含まれており、「ひとり暮らし高齢者」のような高齢者に関する地方自治体が持っている課題が含まれていることも特徴である。

次に、「財源」の分散表現とコサイン類似度が高い単語上位 10 件を、それぞれのモデルで出力した結果を表 3 に示す。 Wikipedia モデルでは、「高齢者」の場合と同様に、「財源」と類似した文脈で出現する別の意味の単語が見られる。地方議会会議録モデルでは、「財源」の細分類を表す単語が多く含まれ

ていることがわかる。

3. ETM を用いたトピック分析

3.1 目的

本節では、東京都議会会議録に対して、 Wikipedia と地方議会会議録から学習した分散表現を用いたトピック分析実験を行う。この実験では以下の 2 点について調査することを目的とする。

- 学習データの違いによって、得られた分散表現がトピック分析にどのような影響があるのか
- トピック数の違いによって、出力するトピックにどのような違いがあるのか

3.2 実験データ

本研究では、NTCIR-14 QA Lab-PoliInfo の Segmentation Task のデータを使用してトピック分析を行う。この Segmentation Task は東京都議会の 2011 年第 3 回定例会から 2012 年第 4 回までの定例会をデータとして使用し、都議会だよりで引用された文に対応する発言範囲を会議録から特定するタスクである。このデータの解答には求める発言範囲に加えてトピックも示されていることから、トピックに対する発言範囲をトピック単位の発言として使用する。トピック分析を行った結果の中に発言範囲中の単語が数多く含まれることが望ましいと考えられる。実験では、Segmentation Task の会議録データから得られた 465 件の発言をトピック単位の発言としてトピック分析で使用する。図 1 に示したように、トピック単位として用いる文書は、あるトピックに対する議員の質問と知事側の答弁を合わせた発言である。

3.3 発言の単語列への変換

トピック単位の発言のすべてに対して、形態素解析器は MeCab を利用して単語に分割する。このとき、辞書には 第 2. 節で使用した Comainu を用いて形態素解析を行う。形態素解析の結果、名詞・動詞・形容詞と判定された単語を取り出すことで、発言を単語列の表現に変換する。

3.4 発言のベクトル化

各発言に対して出現単語を TF-IDF で重み付けし、発言をベクトルで表現する。それぞれの発言における出現単語の頻度と発言全体における出現単語の頻度を計算することで、発言中の単語に対する TF-IDF 値を求めることができる。これにより、それぞれの発言は全単語の TF-IDF 値を要素とするベクトルとして表現される。

トピック	都議会だより(294号)	東京都議会会議録 2011年第3回定例会
新内閣への建言 知事が込めた想いは。 知事 首都の知事として強い危機感に立ち、現場を踏まえて緊急になすべきことを建言した。日本再生に向けて速やかに行動して、都民・国民の不安を振り払ってもらいたい。	質問	東日本大震災から半年が経過いたしました。被災地には、未曾有の大災害のつめ跡が依然として残り、被災された方々の苦難が続いております。東京にも、約八千人の方が避難を余儀なくされております。 ... 地震が起きたから東京の機能を分散させるというのは余りに短絡的で、大都市が国家を引っ張っていくという二十一世紀の常識を無視した暴論であります。常に国家を憂い、日本の将来を見据えて発言し、行動してきた知事に建言に込めた想いを伺います
トピック 「10年後の東京」計画 今回の改定に向け、区市町村からの要望をどう生かしていくのか。 知事本局長 少子高齢社会への取組や防災力強化等、多くの要望を検討し計画に反映する。	答弁	まず、新内閣への建言に込めた想いについてであります、原発事故への対応など、国の政つかつてなく混迷し、国民は先が見えない不安を募らせております。 ... 近々、総理にも直接会うつもりでありますが、今後、日本再生に向けて速やかに行動して、都民・国民の不安を振り払ってもらいたいものだと思います。

図 1: トピック単位として用いる文書の例

3.5 ETMによるトピック分析

トピック単位の発言ベクトルに対して、Embedded Topic Modeling (ETM) [1] を用いてトピックモデルの学習を行う。ETM 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。本稿では、地方議会会議録から得られた分散表現と Wikipedia から得られた分散表現の二種類を用いて、それぞれの分析結果についてトピックに含まれる単語を比較する。ETM で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、トピック数を 50,100,300、トピックの学習回数を 1000 回と設定してトピック分析を行った。

4. 考察

「都議会だより」に付与されているトピックの総数は、文書数と同じ 465 トピックである。そのうち、重複しているものを除くと、異なりトピック数が 283 件となる。さらに、その 283 件を著者の主観によりカテゴリに分けると、下記の 37 カテゴリとなった。「アジア」「医療」「介護・福祉」「外国人・外交」「海洋政策」「整備・対策」「教育・子育て」「環境・エネルギー」「まちづくり」「計画」「経済」「交通」「国政」「自転車」「社会保障」「障害者支援」「職員・職場」「スポーツ」「税金」「地域」「知的財産・情報管理」「中小企業」「選挙・投票・条例」「都市開発」「都政」「農業」「犯罪」「被災地支援」「文化」「防災」「法律・条例」「保険」「ものづくり」「産業・雇用」「少子・高齢社会」「都政運営」「その他」これらの中でも、「介護・福祉」「教育・子育て」「防災」に関する内容が多く含まれることから、この 3 つの内容に対して「トピック数」および「分散表現」の観点から考察する。

4.1 分散表現

分散表現による出力結果は、表 2 で述べたように、Wikipedia と地方議会会議録の学習するコーパスにより異なる。ここでは、3 つのトピックに焦点を当て、分散表現の違いがどの程度異なるのかを明らかにする。

ここでは、上位 30 単語に関連単語が 2 単語以上含まれるトピックを関連トピックとみなし、関連トピック数をカウントする。「福祉・介護」に関するトピックは、50 トピックのうち、地方議会会議録コーパスで 4 トピック、Wikipedia コーパスで 3 トピックあることを確認した。関連する単語は「高齢」「介護」「支援」「福祉」「保険」「施設」とした。「教育・子育て」に関する単語を含むトピックは、50 トピックのうち、関連単語が 2 回以上含まれるトピックが、13 トピックずつあることを確認

した。関連単語は「高校」「教師」「育成」「家庭」「教育」「学校」「人材」「子供」「保育」「発達障害」「児童」「子育て」とした。「防災」に関する単語を含むトピックは、50 トピックのうち、関連単語が 2 回以上含まれるトピックが、地方議会会議録で 10 トピック、Wikipeida で 15 あることを確認した。関連単語は「浸水」「被害」「災害」「対策」「被災」「地震」「減災」「震災」とした。カ「防災」の場合には、地方議会会議録コーパスを用いた方が、トピックを絞り込んでいるが、「福祉・介護」「教育・子育て」についてはトピック数がほぼ変わらない結果となった。

次に、「東日本大震災」という具体的なトピックが含まれる単語集合を比較することとした。上位 30 単語に「東日本大震災」を含むトピックはそれぞれ 3 トピック存在した。下記にその結果を示す。

下記は、「地方議会会議録コーパス」の 3 つのトピックに対して関連単語を用いて筆者が作成した「ラベル」と上位 30 の関連単語である。

- 「東日本大震災」を踏まえた安全の取り組み
「する」「成る」「踏まえる」「思う」「取り組み」「東日本大震災」「果たす」「安全」「協定」「強い」「優れる」「厳しい」「生活」「避難場所」「皆さん」「世界」「規制緩和」「発表する」「聞く」「減少する」「目的」「部分」「人間」「指導」「仕組み」「二十二年」「迎える」「都市計画道路」「可能性」
- 「東日本大震災」社会や中小企業の対策
「対策」「中小企業」「東日本大震災」「高い」「確保」「答える」「不安」「再生」「一つ」「道路」「影響」「社会」「駐車場」「聞く」「申し上げる」「導入」「強い」「対応する」「地震」「見直す」「政府」「除染」「押さえる」「全国」「予定」「見据える」「責任」「教員」「機会」
- 「東日本大震災」取り組みや施策を考える
「考える」「施策」「取り組み」「東日本大震災」「節電」「取り組む」「厳しい」「事業」「成果」「議論」「招致活動」「以上」「大会」「市町村」「御伺い」「関心」「福島県」「成す」「早期」「自分」「感謝」「残る」「御所見」「避難場所」「過去」「働き掛け」「与える」「整備する」「全力」

下記は、「Wikipedia コーパス」の 3 つのトピックに対して関連単語を用いて筆者が作成した「ラベル」と上位 30 の関連単語である。

1. 「東日本大震災」高齢者や介護に対応する
「東日本大震災」「致す」「今回」「強化」「対応する」「現状」「高齢者」「行く」「応ずる」「来る」「七月」「現在」「介護」「変える」「計画」「続く」「引き続く」「調査する」「少ない」「住宅」「共助」「来年度」「認識」「低い」「参加する」「創設する」「作業」「福島県」「中学校」
2. 「東日本大震災」区市町村の被害への支援を進める
「する」「進める」「東日本大震災」「区市町村」「支援」「被害想定」「作る」「自助」「事業」「上げる」「影響」「安全」「生かす」「考え方」「防災対策」「対応する」「実態」「最後」「方々」「取る」「担う」「有する」「行く」「支援策」「比べる」「全国」「地域防災計画」「情報提供」「危険性」
3. 「東日本大震災」安全や支援の取り組みを行う
「行う」「支援」「取り組み」「東日本大震災」「無い」「活用」「公表する」「聞く」「策定する」「安全」「実現する」「開催する」「耐震化」「再生」「拡充」「東京湾」「研修」「育てる」「努める」「生活」「エネルギー政策」「与える」「来る」「条例」「連続立体交差化」「多い」「福島」「果たす」「比べる」

「東日本大震災」については「地方議会会議録コーパス」「Wikipedia コーパス」ともに同じクラスタ数であり、内容についても大きな差をみつけることができないが、長単位解析ツール Comainu を用いたことで関連トピックと発言を結びつけやすい。例えば、平成 23 年第 3 回定例会の増子博樹議員の発言は「地方議会会議録コーパス」の 1 番目のトピックや「Wikipedia コーパス」の 3 番目のトピックに関連していると考えられる。

平成 23 年第 3 回定例会 増子博樹議員の発言

都は、帰宅困難者の一時待機施設の安全性に配慮しながら、都内各所に多くの施設を確保していくべきと考えますが、所見を伺います。東日本大震災発災から二日後、関西広域連合は、現地連絡所を岩手県、宮城県に設置し、また、十六日には福島県に設置して、被災地支援に取り組み始めました。

一方で、MeCab に IPADIC の辞書を用いた場合には「東日本大震災」と一単語とならず「東日本」「大震災」の二単語に分割されることからラベル付けが困難となる。

4.2 トピック数

本実験では、トピック数を 50, 100, 300 と設定した。これらのトピック数は、人手で分割した 37 のカテゴリに近い 50 トピック、カテゴリをさらに細分化した場合を想定した 100 トピック、異なりトピック数 284 件に近い 300 トピックを考えたためである。ここでは、前節の例で用いた「東日本大震災」が含まれるトピック数を調べた。トピック数を 100 にした場合には地方議会会議録で 7 トピック、Wikipeida で 6 トピックに含まれており、トピック数を 300 にした場合には地方議会会議録で 14 トピック、Wikipeida で 12 に含まれていた。トピック数を増やすことで「東日本大震災」という大きなトピックから、「耐震化」や「節電」のようなサブトピックを推定できる可能性がある。しかしながら、上位の関連単語からだけは推定することが困難であり、利用方法について検討の余地がある。

5. おわりに

本稿では、地方議会会議録に対して、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テキストマイニング手法を利用した分析を行った。実験では、NTCIR14 Segmentation task で利用されたデータセット^{*4}を用いて、単語分割、および、学習データの違いにより、トピックモデルの結果がどの程度異なるのかを明らかにした。単語分割については、Comainu を用いたことにより、固有名詞や複合名詞を扱えるようになり、トピックの意味が理解しやすくなり、ラベル付けが容易できることを確認した。学習コーパスについては、地方議会会議録を学習データとした分散表現を用いることで、細かな表現（例えば「高齢者」の類義語の豊富さ）に対応できる可能性があることを確認した。しかしながら、大規模地方議会会議録と Wikipedia のコーパスによる明確な違いは確認できていない。今後は、トピック分析を用いて、異なる議員の発言内容の中から、関連した質問内容をみつけることを考えている。

謝辞

本研究はセコム科学技術振興財団の助成を受けています。

参考文献

- [1] Adji B Dieng, Francisco J R Ruiz, and David M Blei. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*, 2019.
- [2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS '13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.
- [3] 菅原 晃平, 大 城卓, 斎藤 誠, 永井 隆広, 渋木 英潔, 木村 泰知, and 森 辰則. 地方議会会議録コーパスの拡充における問題点の分析と対処. In 言語処理学会第 18 回年次大会発表論文集, pages P1–15, 2012.
- [4] 松河 秀哉, 大山 牧子, 根岸 千悠, 新居 佳子, 岩 千晶, and 堀田 博史. トピックモデルを用いた授業評価アンケートの自由記述の分析. 日本教育工学会論文誌, 41(3):233–244, 2018.
- [5] 小澤 俊介, 内元 清貴, and 伝 康晴. Bccwj に基づく中・長単位解析ツール comainu. In 言語処理学会第 20 回年次大会発表論文集, pages 582–585, 2014.
- [6] 斎藤 誠, 大城 卓, 菅原 晃平, 永井 隆広, 渋木 英潔, and 木村 泰知. 地方議会会議録の収集とコーパスの構築. In 言語処理学会第 17 回年次大会発表論文集, pages 368–371, 2011.
- [7] 木村 泰知, 渋木 英潔, 高丸 圭一, 乙武 北斗, and 森 辰則. 地方議会会議録コーパスの構築とその利用. In 人工知能学会第 26 回全国大会論文集, page 3B3NFC43, 2012.

^{*4} <https://github.com/kmr-y/NTCIR14-QALab-PoliInfo-FormalRunDataset>