市民のツイートを行政課題ごとに分類するための関連語の抽出

南澤　亜樹†　　関　　洋平††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †s1411576@u.tsukuba.ac.jp, ††yohei@slis.tsukuba.ac.jp

あらまし　市民のツイートを分析することで，アンケートでは抽出できない市民の声を得ることができる．このようにして得られた市民の声を行政に反映させるためには，行政が対応する必要がある課題（以下，行政課題とする）ごとに適合するツイートを，分類して整理する必要がある．本研究では，対象となる行政課題の関連語を抽出する手法を提案する．提案手法では，行政課題名と分散表現のベクトルの類似度が高い単語から，関連語候補を抽出する．また，関連語候補の中から，他の行政課題名と類似している単語について，フィルタリングを行う．さらに，複数の都市で同じ行政課題について抽出した関連語の重複を調査することで，都市を横断した関連語と個別の都市に依存した関連語とを区別する．複数の都市の市民の一年分のツイートを用いた実験の結果，提案手法と，相互情報量に基づき関連語を抽出する比較手法との間では，抽出した有用な関連語の件数と精度の平均について，t 検定 (有意水準 5%，片側検定) で有意差を確認した．また，抽出した関連語により，行政課題に適合したツイートが拡張できることを確認した．

キーワード　情報検索，関連語抽出，市民，分散表現

1. はじめに

1.1 本研究の背景

市政において，市民の意見を収集し反映させることは重要である．市民の意見を収集するために自治体は，意見交換会やアンケート等を実施している．しかしそれらの方法は，参加できる住民は限られていることや，実施するためのコストがかかるという問題がある．

本研究では，Twitter(注1) に着目する．Twitter とは，ソーシャルメディアの一つで，最大 140 字で短文を投稿することができるサービスである．誰でも簡単に情報発信ができ，リアルタイム性が高い特徴がある．さらに匿名で利用することができることから，ユーザの率直な意見が含まれるツイートが多くなる傾向がある．また，Twitter には，ユーザ自身の情報，例えば居住地などを登録することができ，その情報を元に，同じ地域に住んでいるユーザを推定することができる．それらのユーザのツイートから情報を抽出することで，市民が話題にしている地域の課題を抽出することができる．

以上の手続きにより，Twitter から情報を抽出することで，市民の率直な意見が得られる．また，Twitter から市民の意見を抽出することで，コストをかけることなく地域の課題を収集することができる．さらに，意見交換会に行く時間がない市民でも，ツイートをしていれば，市民の意見として拾い上げることができる．

一方で，ツイートは誰もが自由に投稿できるため，ツイートの話題は多様である．そのようなツイートを行政に反映させるためには，行政が対応する必要がある課題に適合したツイートを分類して整理する必要がある．本研究では，「教育」，「高齢化」，「交通」，「環境」などの行政が対応する必要がある課題を行政課題とし，行政課題を表す単語を行政課題名とする．

(注1)：https://twitter.com/

1.2 地域の情報を抽出する際の課題

Twitter から地域の情報を抽出する研究には，2.1 節で述べるように，栗原ら [7] や六瀬ら [14] の研究などがある．しかし，これらの研究では，明示的な行政課題名が含まれていないツイートには対応できないという欠点がある．ツイートは短文であることから，必ずツイート中に行政課題名を含むとは限らない．そのため，行政課題名のみでは，行政課題に適合する市民のツイートを十分に分類できない．

例えば，「体育館」に関するツイートを抽出する場合を考える．

- 体育館の予約ってどうするんだろう
- アリーナのトレーニングルームって勝手に使っていいのかな

上記は，「体育館」に関するツイートの例である．「体育館」という行政課題名のみでは，前者のツイートは抽出できるが，後者のツイートは，「体育館」という行政課題名は含まれていないため，抽出できない．本研究では，このように行政課題名が含まれていないツイートにも対応するために，行政課題名に関連する語（関連語）の抽出を行う．

1.3 関連語の抽出における課題

関連語の抽出に関する研究は 2.2 節で述べるように，田中ら [12] や，白木原ら [10]，有賀ら [4] など多くの研究がある．これらの研究では，関連する語を抽出するだけで，関連語が複数のクエリに対して重複しているかの考慮はなされていない．本研究では，同じ部局に関係のある行政課題を区別して市民のツイートを分類することを目的としている．そのためこれらの手法をそのまま用いると，関係がある行政課題間で，関連語候補が重複してしまい，行政課題名に適合したツイートを収集する際にノイズが含まれてしまう．そこで本研究では，関連語候補

の中から複数の行政課題名に関連していない単語を抽出するための手法を提案する．

本研究では，分散表現を用いて関連語の抽出を行う．分散表現を用いることで,「公園」と「広場」のように，類似した単語を取得でき，類似した内容のツイートを合わせて抽出できる．

しかし，分散表現を用いた関連語の抽出には，課題がある．関係のある行政課題について同じ関連語が抽出される場合があるが,「水質汚染」と「大気汚染」のような行政課題においては，対処方法が異なることから，それらを区別する必要がある．そこで本研究では，分散表現を用いて関連語候補とそれぞれの行政課題名に対する類似度を計算し，関連語候補を抽出する．また，関連語候補の中から，他の行政課題名と類似している単語について，フィルタリングを行う．これにより，本研究の目的である，行政課題ごとにツイートを分類するための関連語が抽出できる．

1.4 論文の構成

本稿の構成を以下に示す．2 章では，Twitter から情報を抽出する研究と，関連語の抽出に関する研究を紹介し，本研究の位置づけを述べる．3 章では，本研究の提案手法について詳細を述べる．4 章では，抽出した関連語の有効性と，抽出した関連語を利用したツイートの行政課題に対する適合性を検証することにより，提案手法の評価を行う．最後に 5 章では，本研究で得られた知見をまとめ，今後の課題を述べる．

2. 関 連 研 究

2.1 市民のツイートを抽出する研究

市民が自由に発言をしている Twitter からツイートを抽出することで，アンケート等では得られない率直なツイートが得られる．また，意見を収集する場を設けなくても市民の声が得られるため，コストが削減できるという利点もある．そのため，Twitter から様々なツイートを抽出する研究が行われている．

栗原ら [7] は，自治体名を含んだツイートと，時間的に近接するツイートを対象にし，自治体と強く関連のある名詞を元に作成した辞書を作成し，要望抽出を行った．この研究では，ツイートの発信元を特定の都市に限定していない．本研究では，特定の都市の市民が発信しているツイートに着目して，その都市に関するツイートを抽出する．

六瀬ら [14] は，特定エリアの話題を抽出する研究を行った．具体的には，キーフレーズ API を用いて関連語を抽出し，時間経過によって重要度が減少する重みづけを行うことで，リアルタイム性が高い話題を抽出した．

関 [11] は，ソーシャルメディア上にある市民ツイートを抽出し，市民ツイートが活用できるかと，市民ツイートを市政に反映した場合の，市民ツイートの傾向の変化を検証した．具体的には，自治体が開催したイベント期間中に，自治体名などのキーワードが含まれたツイートを収集し，先行研究により作成した要求表現等が含まれているツイートをポジティブ，ネガティブ，その他のツイートに分類した結果，市民の不満を抽出し，自治体側で不満を解消するための情報発信に活用した．

以上のように Twitter の投稿を利用することで，地域の情報を抽出することができる．そのため，本研究においても Twitter を利用して地域の情報を抽出する．一方で，これらの研究では，市民のツイートを抽出するだけで，行政課題ごとにツイートを分類していない．ツイートを市政に反映させるためには，市民のツイートが行政課題ごとに分類することが有効である．すべての市民のツイート集合から課題を見つけ出し対処法を考えるのは，時間と労力がかかり困難である．しかし，行政課題ごとにツイートが分類されていれば，行政課題ごとに対処法を考えれば良く，市政においてツイートを活用しやすくなる．一方で，ツイートは短文であることから，必ずツイート中に行政課題名を含むとは限らない．そのため，行政課題名のみでは，行政課題に適合する市民のツイートを十分に獲得できない．

本研究では，行政課題ごとに市民のツイートを分類するための関連語の抽出手法を提案する．それぞれの行政課題に特有の関連語を抽出することで，行政課題名を明示的に含まないツイートを拡張し，行政課題ごとにツイートを分類する．

2.2 関連語の抽出に関する研究

関連語の抽出に関する研究は，Twitter やブログなど様々な媒体を対象に行われている．Twitter から関連語の抽出を行った研究には田中ら [12]，白木原ら [10] の研究がある．

田中ら [12] は，ユーザのクラスタに着目し，Twitter からイベントに関する関連語の抽出を行った．検索語についてよくツイートするユーザ群のツイートから，関連語を抽出している．kizAPI(注2) を用いて関連語候補を抽出し，ユーザのクラスタによってスコア付けを行い関連語を抽出した．本研究では，市民に着目してツイートを収集し，分散表現の類似度に基づき関連語を獲得することにより，その都市に特有の行政課題の関連語を獲得する手法を提案する．

白木原ら [10] は，時間的距離に注目して Twitter から関連語を抽出した．関連語の抽出には，RWEA (Related Word Extraction Algorithm) という単語間の距離に着目した重みを，単語の出現頻度に乗じる手法を応用し使用している．また，リアルタイム性の高い関連単語を取得するため，RWEA の単語間の距離を時間的距離に置き換えて使用している．この研究では，リアルタイム性に着目しているが，本研究で獲得する行政課題の関連語は，分散表現を構築する際の収集時期に依存しており，また分散表現を構築するためにある程度の量のツイートを収集する必要がある．以上を踏まえ，リアルタイムに関連語を抽出することは考えず，過去の一定期間に収集したツイートを利用して抽出した関連語に基づき，行政課題に適合したツイートを分類する．

また，Web 上において関連語の抽出を行った研究には，有賀ら [4] や城光ら [9] の研究がある．有賀ら [4] は，文脈に応じた関連語の抽出を行った．英語の Wikipedia を使用し，word2vec [3] で用いられている Continuous Bag-of-Words（CBOW）に語順情報を追加したモデルを用い，文章の中心単語を予測して関連語の抽出を行った．この研究では，関連語の抽出に，文脈や語順情報を利用して中心単語を予測し使用している．城光ら [9]

(注2)：http://kizasi.jp/tool/kizapi.html

は，Skip-gram モデルで関連語の抽出を行った．Skip-gram に周辺単語の品詞，周辺単語との距離の情報を追加したモデルを用いて，単語間の類似度を計算して関連語を抽出している．この研究のように，関連語の抽出には，周辺単語との距離を考慮する方法も考えられる．本研究で対象とするツイートは，短文であることや，くだけた文が多いことから，語順や周辺単語の距離については扱わず，今後の検討課題とする．

これらの研究では，抽出した関連語が曖昧性を持っている，すなわち複数の入力（イベント名や検索語）に対して同じ関連語を抽出していることについては，考慮されていない．本研究では，関係のある行政課題ごとに市民のツイートを分類することを目的としている．そのため，行政課題間に関係がある場合，それぞれの行政課題から抽出した関連語が重複してしまうと，ツイートを分類することが難しくなる．本研究では，関連語の抽出の際に，行政課題間の関連語候補の重複を解消するためのフィルタリングなどの工夫を行う．

3. 提案：行政課題ごとにツイートを分類するための関連語の抽出手法

3.1 提案手法の概要

本章では，市民のツイートを行政課題ごとに分類するための関連語の抽出方法について説明する．本研究における関連語の抽出は，以下の手順で行う．

（1） 関連語候補として，行政課題名と類似している単語を，分散表現を利用して抽出．

（2） 手順 (1) で抽出した関連語候補と類似している上位 n 件の単語を，分散表現を利用して抽出し，以下の条件にあてはまる単語をフィルタリング．

(a) 関係のある行政課題名の中で，対象となる行政課題名が最上位でない．

(b) 対象となる行政課題名への類似度と，他の行政課題名への類似度との差が小さい．

(c) 対象となる行政課題名への類似度と，関係のある行政課題名への類似度がすべて閾値以上．

（3） 都市間で関連語の比較を行い，都市に依存しない関連語，都市に依存する関連語を判別．

また，提案手法の全体図を図 1 に示す．

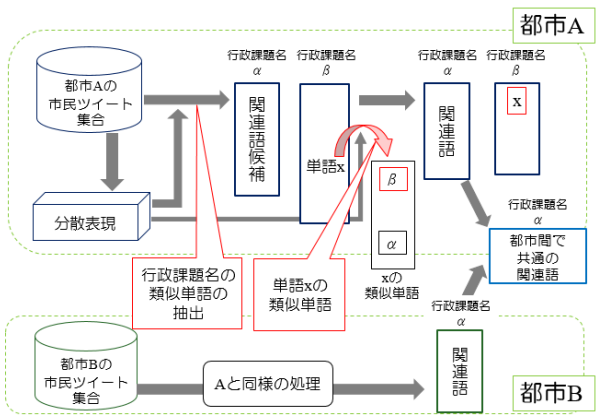

図 1 提案手法の全体図

3.2 関連語候補の抽出

本研究では，行政課題ごとに適合する市民のツイートを分類することを目的とする．ここで行政課題とは，行政が対応すべき課題を指し，本研究では，「教育」，「高齢化」，「交通」などに関連する課題のことを表し，行政課題を表す単語を行政課題名とする．行政課題に適合する市民のツイートを収集するには，行政課題名による収集が考えられる．しかし，単純な行政課題名による収集では，行政課題名を明記しているツイートしか収集することができず，充分な数の行政課題に関連する市民のツイートを分類できない．そこで本研究では，市民のツイートを行政課題ごとに分類するための関連語を抽出する．

関連語の抽出には，共起頻度や，相互情報量 [1] を用いる手法がある．しかし，頻出単語や相互情報量には，課題名を含んだテキストを用いて抽出を行うため，同じテキスト中に課題名と共起した単語しか抽出することができないという課題がある．ツイートは短文であることから，行政課題名に似た意味を持つ単語が，行政課題名と同一のツイートに含まれているとは限らない．本研究では同一ツイート中に明記されていない関連語も抽出するために，周辺の単語を予測できるように対象となる単語のベクトルを構築する Skip-gram に基づく分散表現を利用する．単語の類似度は，分散表現に基づくベクトル間のコサイン類似度を用いる．分散表現は，宮尾ら [13] によると「単語やフレーズ，文の意味がすべての実数ベクトルで表現されるため，単語や文の近さはすべてベクトルの近さとして計算できる」ため，単語の類似度を計算することができる．

3.3 関連語候補のフィルタリング

類似度の上位の単語をそのまま用いて関連語を抽出すると，対象となる行政課題に関係のある行政課題が存在する場合，関連語候補が重複する可能性がある．この場合，その単語を用いてツイートを拡張すると，行政課題に直接関係のないツイートも含まれてしまう．

これを解消するために本研究では，行政課題名の関連語候補と類似している上位 n 件の単語を抽出し，関係のある行政課題名の中で，対象となる行政課題名が最上位でない単語をフィルタリングをする．また，行政課題名への類似度と，他の行政課題名への類似度との差が小さい単語は，関連語候補からフィルタリングをする．さらに，関係のある行政課題全般に対する類似度が高い単語は，一般的なより広い意味で用いられている単語とみなし，フィルタリングをする．

3.4 都市間で共通の関連語と都市独自の関連語の区別

それぞれの行政課題の関連語を，複数の都市で抽出し比較することで，どの都市にも共通して現れる行政課題の関連語と，ある都市特有の行政課題の関連語とを区別して抽出する．

4. 実験：関連語により拡張したツイートの行政課題に対する適合性判定

4.1 実験の目的

本実験では，市民のツイートを行政課題ごとに分類するために，(1) 行政課題ごとに関連語が抽出できたか，(2) 抽出した関連語により行政課題に適合するツイートを拡張できたかの 2

点について検証を行う．

4.2 使用するデータ

使用するデータは，2017 年 1 月から 2017 年 12 月の 1 年間に，水戸市民と，つくば市民について，プロフィールに水戸やつくばと明記されたユーザが投稿したツイートを対象とする．ツイートの収集には，Twitter の Streaming API(注3) を利用した．水戸市民やつくば市民のアカウントは，ツイプロにその都市の市民として登録されているユーザとそのフォロワーを対象として，水戸やつくばとプロフィールに明記されているユーザを選別している．リツイートや bot が行ったと推定されるツイートおよび重複するツイートは削除する．以上の手順により得られたアカウント件数とツイート件数を，表 1 に示す．

表 1　対象データのサイズ

	水戸市	つくば市
アカウント件数	7,291	9,987
ツイート件数	3,168,455	7,617,809

4.3 実験方法

(1) 水戸市民とつくば市民のツイートを用いて,「保育園」,「幼稚園」に対する行政課題ごとの関連語の抽出と,「バス」,「電車」に対する行政課題ごとの関連語の抽出に関する実験を行う．関連語の抽出には，提案手法と相互情報量を用いた比較手法を採用し，同じ行政課題について，水戸市とつくば市との両方で抽出された関連語について評価する．さらに，それぞれの都市だけで抽出された関連語について考察する．

(2) (1) で抽出した関連語のなかで，有用と第一著者が判断した関連語のうち，類似度が高い単語を人手で最大 5 件選択し，選択した関連語を含むツイートを 5 件ずつランダムサンプリングで抽出し，行政課題名に適合しているか，第一著者が評価を行い，その精度を，提案手法について評価する．

- 提案手法は，3 章で述べたように，分散表現を用いて関連語を抽出する．それぞれの市民のツイートに対して構築された分散表現のモデルを用いて,「保育園」,「幼稚園」,「バス」,「電車」に類似している単語を上位 200 件まで抽出する．

以上の手順により抽出された関連語候補それぞれに対して，分散表現に基づき類似している単語，上位 n 件を抽出する（本実験では，$n = 50$ とする.）．その後，行政課題ごとの関連語とするために，以下のフィルタリングを行う．なお，以下のパラメータは，著者が経験に基づき設定した．

(a)　関係のある行政課題間の中で，対象となる行政課題名が最上位でない場合，関連語候補からフィルタリングする．

(b)　関連語候補の各行政課題名に対するベクトル類似度を抽出し，対象となる行政課題名への類似度と他の行政課題名への類似度との差が閾値 α 以下の単語を，関連語候補からフィルタリングする.（本実験では，$\alpha = 0.02$ とする.）.

(c)　対象となる行政課題名への類似度と，類似した他の行政課題名への類似度すべてが閾値 β 以上の単語を，関連語候補からフィルタリングする.（本実験では，$\beta = 0.5$ とする.）.

(注3)：https://developer.twitter.com/en/docs

本実験では，関係のある行政課題の組を「保育園」,「幼稚園」,「小学校」と「バス」,「電車」としている.「保育園」と「幼稚園」の関連語の中には,「保育園」や「幼稚園」よりも「小学校」に関連している単語が含まれてしまうことから，行政課題として「小学校」を追加して実験を行う (注4).

また，行政課題名が含まれるツイート数と関連語候補が含まれるツイート数を調査し，行政課題名が含まれるツイート数の半数よりも多い関連語は，行政課題に関連のないツイートが多いと考えられるため，その関連語候補についてもフィルタリングを行う．さらに，抽出された関連語候補の中には，意味を持たない単語（例えば,「ゃないけどさぁ」）が含まれることから，品詞でフィルタリングを行う．品詞のフィルタリングでは，関連語候補を形態素解析器 MeCab [2] を用いて品詞を解析し，品詞が記号になる単語は削除した．また，名詞の中で関連のない単語が多いと予測される,「名詞，一般」で文字列すべてがひらがなの名詞や，動詞や文字数が 1 文字のみの単語は削除した．

- 比較手法は，先行研究 [6] でも用いられている相互情報量 [1] を用いる．比較手法では，提案手法と同様に,「保育園」,「幼稚園」,「小学校」と「バス」,「電車」について相互情報量を算出し，値の比較を行う．相互情報量は，各行政課題に偏った単語が上位にくる指標であり，以下の式（1）で表す．

$$I(U;C) = \frac{N_{11}}{N} log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \quad (1)$$

カテゴリ A の単語 w の相互情報量を求めるとすると，式（1）の，N_{11} はカテゴリ A に単語 w が存在する件数を表す．N_{10} はカテゴリ A に単語 w が存在しない件数を表す．N_{01} はカテゴリ A でないカテゴリに，単語 w が存在する件数を表す．N_{00} はカテゴリ A でないカテゴリに単語 w が存在しない件数を表す．なお，$N_{1.}$ は，$N_{10} + N_{11}$ を示す．

4.4 実験環境

本節では，分散表現の構築について説明する．分散表現の構築には，まず，ツイートを，形態素解析器 MeCab [2] を用いて基本形で分かち書きを行う．使用する単語辞書は，NEologd [8] を用いる．分かち書きされたツイートに対して，word2vec [3] を用いて分散表現を構築する．word2vec の実装では，Skip-gram を使用し，階層的ソフトマックスを使用した．ベクトルに表すための次元と窓幅は，100,200,300,500,1000 次元で，窓幅は 5,10,15,30,50,100 を組み合わせた実験を行い，人手で結果を確認し，最適と思われる 200 次元，窓幅 30 で実験を行う．

4.5 実験結果

(1) 提案手法を用いて抽出された関連語と，比較手法を用いて抽出された関連語を確認する．まず，提案手法，比較手法を用いて水戸市とつくば市で抽出した関連語のうち，行政課題ご

(注4)：例えば,「保育園」の類似している単語に,「学童」のように「保育園」よりも「小学校」に関連する単語が抽出されてしまう．これは,「保育園」と「学童」が子どもを預ける場所という観点から類似しているため抽出されたと考えられる．

とにツイートを拡張する関連語として第一著者が有用と判断した関連語の件数と精度を，表 2，表 3 に示す．また，提案手法と比較手法を用いて抽出された関連語のうち，水戸市とつくば市で共通に獲得できた関連語の上位 5 件を，提案手法については，表 4 に，比較手法については，表 5 に示す．

表 2　二都市で共通している関連語のうち有用と判断した関連語の件数

行政課題名	提案手法	比較手法
保育園	9	4
幼稚園	5	4
バス	13	5
電車	11	6
平均	9.50*	4.75

* t 検定（片側検定，有意差水準 5%，$p = 0.022$）で有意に向上

表 3　二都市で共通している関連語のうち有用と判断した関連語の精度

行政課題名	提案手法	比較手法
保育園	0.90	0.36
幼稚園	0.36	0.17
バス	0.87	0.20
電車	0.73	0.38
平均	0.71*	0.27

* t 検定（片側検定，有意差水準 5%，$p = 0.013$）で有意に向上

表 4　二都市で共通して抽出された関連語上位 5 件（提案手法）

順位	保育園	幼稚園	バス	電車
1	保育	幼児	バス停	新幹線
2	保育所	年少	乗り場	満員電車
3	保育士	親御	路線バス	始発駅
4	送り迎え	才子	シャトルバス	ラッシュ時
5	ママ友	習い事	停留所	地下鉄

表 5　二都市で共通して抽出された関連語上位 5 件（比較手法）

順位	保育園	幼稚園	バス	電車
1	保育	園児	高速	車内
2	落ち	学校	高速バス	満員電車
3	日本死ね	小学	水戸	満員
4	保育園落ちた日本死ね	園生	茨城	遅れ
5	お迎え	子供	路線	通勤

表 2，表 3 の結果より，提案手法で抽出した有用な関連語の件数と精度が，全ての行政課題において比較手法を上回ることを確認した．また，提案手法で抽出した有用な関連語の件数と精度の平均は，比較手法に対して，対応のある片側 t 検定（有意水準 5%）で有意差を確認した．

表 4 をみると，都市に依存しない関連語は，おおむね各行政課題に関連する単語が抽出できた．特に「保育園」の関連語では，7 位に「保活」という単語が抽出できた.「保活」とは，保護者が子どもを保育園に入れるための活動であり,「保育園」に関連のある単語といえる．一方で，表 3 における「幼稚園」の有用な関連語の精度は低かったが，これは「親御」や「習い事」といった，子どもに関係する行政課題全般に関連する単語を削除しきれなかったためである．

表 5 をみると,「保育園」の関連語の 4 位に「保育園落ちた日本死ね」が,「幼稚園」の 21 位に「森友（学園）」が出現しており，比較手法では，特定の時期に話題になった関連語が抽出される傾向がある．また，比較手法では,「幼稚園」の 5 位に抽出された「子供」や,「バス」の 12 位に抽出された「運転」のように，一般的な単語などノイズが多く抽出された．一方で提案手法では，他の行政課題名と類似している単語のフィルタリングを行ったことでノイズが少なくなっており，提案手法の有効性を確認した．

また，提案手法において，水戸市，つくば市それぞれで抽出された関連語から，二都市に共通な関連語を除いた結果をみると,「バス」の関連語では，水戸市では「茨城交通」(注5) や,「いばっピ」(注6) という，茨城交通のバスで使用できる IC カード名が抽出された．茨城交通は水戸市内を中心に茨城県北部で運行しており，水戸市の「バス」特有の関連語といえる．同じくバスの関連語で，つくば市では「つくば号」(注7) という，つくば ～ 東京駅を結ぶバス名が抽出され，つくば市の「バス」特有の関連語といえる．

(2) 提案手法で抽出した関連語の中で，有用と第一著者が判断した関連語のうち，類似度の高いものを人手で最大 5 件抽出する．各行政課題について，提案手法により抽出した関連語により拡張したツイートのうち，ランダムサンプリングで抽出した 5 件を用いて行政課題に対して適合性判定を評価した際の精度の平均を，表 6 に示す．また，各行政課題について，それぞれの関連語により拡張されたツイートの件数と適合性を判定した精度の評価結果を,「保育園」は，表 7 に,「幼稚園」は，表 8 に,「バス」は，表 9 に,「電車」は，表 10 にそれぞれ示す．

表 6　提案手法で抽出した関連語により拡張したツイートの精度の平均

	保育園	幼稚園	バス	電車	平均
提案手法	0.80	0.27	0.80	0.90	0.69

表 7　提案手法で抽出された関連語で拡張したツイート件数と精度（保育園）

関連語	保育		保育士		保育所		慣らし保育		保活	
自治体名	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市
件数	257	487	252	449	96	169	11	21	4	10
精度	0.6	0.8	1.0	0.4	0.8	0.4	1.0	1.0	1.0	1.0

表 6 から,「保育園」,「バス」,「電車」の関連語として二都市で抽出された単語を用いて拡張したツイートは，ランダムサンプリングによる評価の結果，おおむね各行政課題について言及しているツイートであることが確認された．しかし,「幼稚園」

(注5)：http://www.ibako.co.jp/
(注6)：http://www.ibako.co.jp/regular/ibappi/
(注7)：http://kantetsu.co.jp/bus/highway_tsukuba_tokyo.html

表 8　提案手法で抽出された関連語で拡張したツイート件数と精度（幼稚園）

関連語	年少		年長		園生		園長先生	
自治体名	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市
件数	29	55	89	185	12	18	19	48
精度	0.2	0.2	0.2	0.2	0.6	0.4	0.4	0.0

表 9　提案手法で抽出された関連語で拡張したツイート件数と精度（バス）

関連語	バス停		路線バス		シャトルバス		停留所		乗り場	
自治体名	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市
件数	613	1,430	366	406	273	317	57	161	77	101
精度	0.8	0.8	1.0	1.0	1.0	0.8	0.8	0.8	0.6	0.4

表 10　提案手法で抽出された関連語で拡張したツイート件数と精度（電車）

関連語	満員電車		通勤ラッシュ		終電		通勤快速		駅員	
自治体名	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市	水戸市	つくば市
件数	331	994	60	200	921	2,747	45	165	247	622
精度	1.0	1.0	1.0	1.0	0.8	0.8	0.8	1.0	0.8	0.8

の関連語を用いた評価については，精度が低くなっている．

次に，提案手法により抽出された関連語で拡張されたツイートについて検証する．「保育園」の関連語で特に有用と考えられる「保活」で抽出されたツイートの例を以下に示す．

- 田舎だったら待機児童はいないなんて嘘．保活しているママに田舎に引っ越せば待機児童にならないと言った人は考えを改めて欲しい．茨城県でも待機児童率は高い．

このように，ツイート中には「保育園」という文字列が含まれていないツイートも，関連語を用いることで拡張できることを確認した．同様に，「電車」の関連語である「通勤ラッシュ」でも，「電車」という文字列を含まない関連語を抽出でき，行政課題名を含まないツイートを拡張できる．また，これらの単語は，提案手法でしか抽出できないことから，提案手法の有効性を確認できた．

一方で，表 8 の結果のように，「幼稚園」の関連語で拡張されたツイートの精度が低くなっている．これは「年少」「年長」が，「幼稚園」の年齢分け以外にも用いられていることが原因と考えられる．

5. おわりに

本論文では，二都市に共通する関連語の抽出実験を通して，分散表現を利用する提案手法は，比較手法を有意に上回ることを確認した．また，「保育園」の関連語である「保活」などは，行政課題名の文字列が含まれておらず，特に有用な関連語といえる．さらにこれらの単語は，提案手法でしか抽出できないことから，提案手法の有効性を確認できた．

また，都市独自の関連語の抽出結果では，都市特有の関連語を抽出できることを確認した．この結果から，複数の都市で関連語を抽出し比較することにより，ある都市に特有の行政課題に適合するツイートを判別できるといえる．

さらに，提案手法で抽出した関連語を利用して拡張したツイートの適合性の評価の結果より，多くの関連語について，行政課題に適合するツイートを拡張できることを確認した．行政単語名のみでは獲得が難しいツイートも拡張することができ，提案手法の有効性を確認できた．

今後の課題として，市民の意見を収集する研究の知見 [5] を踏まえつつ，行政課題に反映しやすい市民意見を抽出することを計画している．また，類似した行政課題を自動判別することにより，より多くの行政課題を対象とすることで，関連語の精度を高めていくことを検討している．

謝辞　本研究の一部は，科学研究費補助金基盤研究 B（課題番号 16H02913）の助成を受けて遂行された．

文　献

[1] Christopher, D. Manning; Prabhakar, Raghavan; Hinrich, Schutze; Introduction to Information Retrieval. Cambridge University Press, 506p.

[2] Kudo, Taku; Yamamoto, Kaoru; Matsumoto, Yuji. “Applying Conditional Random Fields to Japanese Morphological Analysis”. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004), Barcelona, Spain, 2004-07, Association for Computational Linguistics, 2004, p. 230-237.

[3] Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. “Efficient Estimation of Word Representations in Vector Space”. Proceedings of the International Conference on Learning Representations (ICLR2013). Scottsdale, Arizona, USA, 2013.

[4] 有賀竣哉, 鶴岡慶雅. “単語のベクトル表現による文脈に応じた単語の同義語拡張”. 言語処理学会第 21 回年次大会発表論文集. 2015, p.752-755.

[5] 柏野和佳子, 平本智弥, 関洋平. “市民意見の収集システムで得られたツイートからの道路・交通に関する意見抽出”. 人工知能学会第 57 回ことば工学研究会. 2018.

[6] 河内沙織, 豊田哲也, 延原肇. “Wikipedia カテゴリおよび自己相互情報量に基づく関連検索キーワード生成による知識拡充支援”. 情報処理学会第 73 回全国大会講演論文集. 2011, p.607-608.

[7] 栗原理聡, 佐々木彬, 松田耕史, 岡崎直観, 乾健太郎. “Twitter を利用した地域毎の要望抽出”. 人工知能学会第 29 回全国大会論文集. 2015, 1H3-3.

[8] 佐藤敏紀, 橋本泰一, 奥村学. “単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討”. 言語処理学会第 23 回年次大会, 言語処理学会, 2017, NLP2017-B6-1. p.875-878.

[9] 城光英彰, 松田源立, 山口和紀. 文脈限定 Skip-gram による同義語獲得. 自然言語処理. 2017, 24(2), p.187-204.

[10] 白木原渉, 大石哲也, 長谷川隆三, 藤田博, 越村三幸. “時間的距離に注目した Twitter からの関連単語抽出”. 情報処理学会研究報告. 2012, vol.2012-NL-205, no.14, p.1-7.

[11] 関洋平. “ソーシャルメディア上の市民意見を利用した市民共創知の可視化”. 市民共創知研究会. 市民共創知研究会資料. 2016, vol.1, no.17, p.1-4.

[12] 田中匠, 関洋平. “マイクロブログユーザのクラスタに着目したイベント手がかり語の抽出”. 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014). 2014, B6-2.

[13] 宮尾祐介. 統計的自然言語処理-ことばを扱う機械：自然言語の意味に対する 2 つのアプローチ. 岩波データサイエンス. 2016, vol.2, p.63-74.

[14] 六瀬聡宏, 清水真, 古橋慎之介, 高畑洋貴, 近藤直人, 佐藤智貴, 遠藤岳, 渡辺雅史, 内田理. “Twitter を用いた特定エリアにおける注目話題の抽出とその可視化”. 電子情報通信学会言語理解とコミュニケーション研究会. 電子情報通信学会. 2014, NLC2013-49, p.11-14.