1H3-3

Twitterを利用した地域毎の要望抽出

Extracting Local Resident Demands per Region Using Twitter

栗原 理聡 *1 Masatoshi KURIHARA
佐々木 彬 *1 Akira SASAKI
松田 耕史 *1 Koji MATSUDA
岡崎 直観 *1 Naoaki OKAZAKI
乾 健太郎 *1 Kentaro INUI

*1東北大学
Tohoku University

For the growth and development of a city, it is essential to hear the voice and opinion of its residents. However, most people who participate in town hall meetings are mainly senior citizens. On the other hand, due to the anonymity that social networking services such as Twitter offer, many young people voice their opinions about a city using the Internet. Therefore, in this work, we propose a Twitter-based system for extracting the realistic demands that local governments are able to handle.

1. はじめに

自治体のまちづくりには住民の声を反映させることが求められる．そのため自治体は，意見交換会やワークショップといった住民との話し合いの場を設け，住民主体のまちづくりとなるように努力を重ねている．しかし，そのような意見交換会やワークショップに住民全員が参加することはほぼ不可能であり，種々の調査によると，参加する人々の多くは 60 歳以上の高齢者であるのが現状である [*1][*2][*3]．そのため，様々な年齢層の意見を集めるのに適した場とはなっていない．この問題に対する改善策として，調査員がその地域を訪れ，インタビューやアンケート調査を行うという手法も模索されているものの，調査員や被験者への負担やコストといった問題が残る．

そこで本研究では，Twitter を用いた自治体への要望抽出手法を提案する．Twitter とは 140 文字以内で文章を投稿するソーシャルネットワーキングサービス (SNS) であり，Twitter には日々ユーザからの日常の出来事が大量に投稿されている．Twitter は基本的に匿名で利用され，それゆえユーザの本音が多く投稿される傾向にあるのも特徴のひとつである．さらに Twitter のユーザの年齢層を見ると，10 代から 20 代といった年齢層に多く利用されていることがわかる [*4]．こういった Twitter の特徴をふまえると，Twitter から若い年齢層の人々の自治体に対する正直な要望を抽出し，それを自治体が参考にすることで，幅広い年齢層の意見を反映したまちづくりを行うことが可能になると期待できる．

しかしながら，Twitter の投稿から自治体に対する要望を抽出するにあたっては，自治体名を含んでいないが，暗に言及するツイートをいかに抽出するかが大きな問題となる．Twitter には 1 投稿あたり 140 文字以下という文字数の制約があるため，ひとつの話題が複数のツイートに分けて記述されることが多々ある．例えば，以下の (1) のツイートの後に同じユーザにより (2) のツイートが投稿された場合，(2) は仙台市に関連するツイートであると考えるのが自然である．

(1) 仙台市なう
(2) 地下鉄で Suica 使えないとかありえない

しかしながら,「仙台市」という自治体名のみでツイートを収集すると (2) のツイートは見逃されてしまう．よって,「自治体名を含まないが自治体に関連するツイート」を考慮することが必要となる．

そこで本研究では，(1) 自治体名を含んだツイートにたいして**時間的に近接するツイートからも要望を抽出する**こと，(2) 自治体と強く関連する語句 (**自治体管理名詞句**) からなる辞書をパターンマイニングによって構築し，これを要望抽出に用いることを提案する．

2. 関連研究

Twitter を利用して地域性のあるイベントや特定の地域の特徴を分析する研究は盛んに行われている．土屋らは路線名が含まれるツイートを機械学習を用いて解析し，鉄道の運行トラブルを抽出する手法を提案した [土屋 13]．山本らは Twitter を用いて生活に関連する単語からなる辞書を作成し，特定の地域の生活情報を抽出する手法を提案した [山本 12]．渡辺らは位置情報を持たないツイートに対して，ある場所に特徴的な建物の名前を含んだツイートから場所を推定し，特定の場所のイベントを抽出する手法を提案した [渡辺 11]．Boettcher らはある地理座標において平常時よりも多く用いられている単語から，その場所でのイベントを抽出する手法を提案した [Boettcher 12]．これらの研究は，ある特定の時間，場所で起きたイベントを抽出し，地域の実情を把握しようとする手法であり，自治体に対する要求など，より直接的な表現に着目することで，まちづくりに活かすために有益な意見を収集できるものと期待できる．

一方，Twitter ではなく，アンケートの自由記述欄から意見や要望を抽出する研究も行われている．永野らはアンケートの自由記述欄から得られたテキストデータに対し，形態素解析や共起ネットワーク分析を行うことにより，多く出てくる単語や単語同士のつながりを見ることで評価の傾向や意見の特徴を把握するという手法を提案した [永野 12]．山本らはアンケートの自由回答欄から要望を抽出する手法として，自由回答の記述の何文目に要望が書かれる傾向にあるのかを分析し，要望文を自動抽出する手法を提案した [山本 06]．大塚らは自由回答

連絡先: 栗原理聡，東北大学工学部情報知能システム総合学科，宮城県仙台市青葉区荒巻字青葉 6-3-09，022-795-7091，022-795-4285，masatoshi_kurihara@shino.ecei.tohoku.ac.jp

*1 http://www.hocacon.jp/image/5bukai/machi/26.5.2_2.pdf
*2 http://www.city.saku.nagano.jp/cms/html/entry/9009/file291.pdf
*3 http://www.city.chino.lg.jp/www/contents/1396942171687/files/anketo.pdf
*4 http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2014/h25mediariyou_1sokuhou.pdf

アンケートにおいて間接的な要求を抽出するための基準として「〜てほしい」に言い換え可能か否かという基準を提案し，機械学習手法により要求を抽出した [大塚 04]．アンケートの自由記述欄を対象とした上記の要望抽出手法の場合，分析対象となるテキストはアンケートの調査実施者に対するものであることから，調査目的にそぐわない内容は記述されにくいという特徴を持つため，「要望であるか否か」という点のみに着目して抽出を行うことが可能であった．しかしながら，Twitter を対象とする場合には，1. 節で述べたように要望の対象が自治体であるかどうかを判別する必要がある．

3. 本研究で抽出する要望の定義

本研究で扱う要望の定義は，大塚ら [大塚 04] によるものを参考にした．大塚らは，要望を「直接要求」と「要求意図」に分け，自由回答アンケートからの抽出を試みていた．一方，Twitter の場合は自由回答アンケートと異なりユーザの独り言や愚痴が投稿される傾向がある．この点に着目し，本研究では要望を「直接要求」「要求意図」「不満」に更に細分化した．

3.1 直接要求

「〜てほしい」「〜てください」「〜てくれ」といった，日本語母語話者のほとんどが「要求」と判断できる表現を**直接要求表現**とし，この表現を含むテキストを**直接要求**と定義する．直接要求の例を以下に示す．

(3) 愛知県の公立高校の入試はどうして、11 日のから一週間で、結果発表も 21 日と遅いのだろう？せめて、もう一週間早くしてほしい。

(4) 神戸市営地下鉄〜IC 連絡定期を発売してくれ〜

(5) 京都市営地下鉄も ICOCA 定期券に対応してください。

3.2 要求意図

「〜べき」「〜がベストだと思う」「〜が必要」といった，「〜てほしい」に言い換え可能な表現を**要求意図表現**とし，この表現を含むテキストを**要求意図**と定義する．要求意図の例を以下に示す．

(6) 京都市は西大路を南北に移動できる交通をもっと整備すべき。

(7) 横浜市交通局は毎年アニメタイアップやってるんだから来年はうたプリでやるといいと思う

(6),(7) における「整備すべき」「やるといいと思う」という表現が要求意図表現である．

3.3 不満

以下の例において示す通り，直接要求や要求意図に当てはまらないテキストであっても，その内容が要望の動機になることがある．

(8) 横浜市営地下鉄の始発遅い、最悪

(9) 市役所の対応悪いわ

(8) は「横浜市営地下鉄の始発を早くしてほしい」，(9) は「市役所の対応を良くしてほしい」という要望を潜在的に含んでいると解釈できる．本研究では，このような要望の動機になる否定的なテキストを**不満**と定義する．

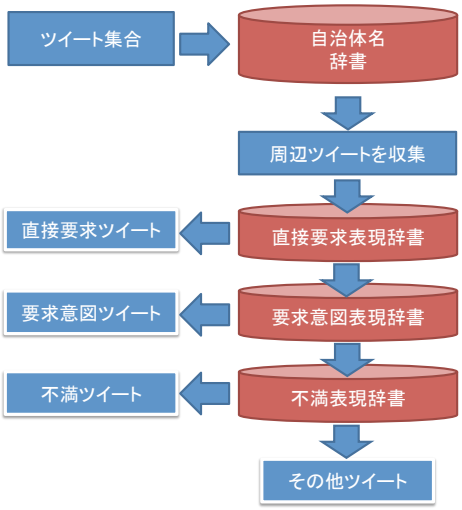

図 1: 直接要求，要求意図，不満抽出の概観

4. 手法

本研究で要望抽出対象とする自治体は，全国 20 の政令指定都市である．また，後述する評価用データとは別に，辞書やルールの開発用データとして 2013 年 2 月，3 月のツイートを用いた．

本研究における提案手法の枠組みを図 1 に示す．まず，自治体名を含むツイートを収集し，さらにそれらの周辺一定時間以内に投稿されたツイートを同様に収集する．次に，収集された各ツイートに対して直接要求表現辞書を用いて，直接要求ツイートを抽出する．その後，要求意図表現辞書，不満表現辞書を用いて，直接要求ツイートでないと判断されたツイートから要求表現ツイート，不満ツイートを同様に抽出する．

4.1 ツイートの収集

自治体に対するより多くの要望を抽出するために，本研究では以下の 2 段階に分けてツイートの収集を行う．

4.1.1 自治体名を含むツイートの収集

はじめに，自治体名をテキスト中に含むツイートを収集する．この際，例えば「福岡」で抽出するのではなく，「福岡市」というように “市” まで含むものに限って収集する．これは，“市” を含まないものを収集してしまうと，「福岡県」に関連するツイートとの区別がつかなくなってしまうためである．

4.1.2 自治体名を含むツイートの周辺一定時間以内に投稿されたツイートの収集

Twitter ではツイートあたりの文字数制約の都合で，ひとつの話題が複数のツイートに分けて記述されることがある．この性質を考慮するために，本研究では 4.1.1 節で収集したツイートと同じユーザにより投稿された，周辺一定時間以内のツイートの収集も行う．

4.2 自治体管理名詞句リストの構築

地域住民の要望を聞くために行われるアンケートの場合，回答者の要望の対象は自治体に対するものとはっきりしているが，ツイートの場合は要望の対象は様々であり，書かれている要望が自治体に関連するものか否かを判別する必要がある．そこで本研究では，自治体にとって対処可能な要望を抽出しやすくするために，自治体の管理対象を表す名詞句を**自治体管理名**

表 1: 獲得した自治体管理名詞句の一部

カテゴリ	自治体管理名詞句
施設名	市役所，図書館，警察署，避難所
インフラ設備名	電気，水道，ガス，地下鉄
役職名	市長，役人，公務員，選管
その他	ゴミ袋，住民税，レンタルサイクル

詞句と定義し，要望抽出の際に利用する．自治体管理名詞句のリストを作成するために，次の手順で作業を行った．

1. 開発用データのツイートから,「(自治体名) の○○」というパターンで語句を抽出する．
2. 語句を出現頻度の降順でソートし，上位 200 件程度を人手でそれぞれ見て，適切でない表現を排除する．

以上の手順により，107 件の自治体管理名詞句を獲得した．獲得した自治体管理名詞句の一部を表 1 に示す．以降の節では，自治体名自体も自治体管理名詞句として扱う．

4.3 要望抽出のためのルール作成

4.2 節で定義した「直接要求」「要求意図」「不満」の各々に対して，抽出するためのルールを作成する．

「直接要求」を抽出するための直接要求表現,「要求意図」を抽出するための要求意図表現の収集にあたっては，大塚らの論文中に記述されている表現リストを参考にした．ただし，大塚らによる表現リストは自由回答アンケートの記述から機械学習手法により分類した直接要求，要求意図の文中に現れた表現のリストであり，自由回答アンケート特有の表現が含まれる．また，Twitter には，自由回答アンケートには見られない,「〜しろや」「〜だろが」といった強い口調の表現も存在する．よって本研究では，開発用データのツイートに対して大塚らの直接要求表現，要求意図表現リストを適用し，Twitter ドメインでは出現しない表現の除去，Twitter ドメインのみに出現する表現の追加を行った．

次に,「不満」を抽出するための手法について説明する．不満の場合，直接要求や要求意図とは異なり，文末表現のみからなるとして抽出ルールを定めることは難しい．3.3 節で述べたように，不満は要望の動機となる否定的なテキストであるということを考慮すると,「自治体あるいは自治体管理名詞句に対して否定的な表現がなされている場合，自治体に対する不満である」と見なすことができると考えられる．よって本研究では，否定的な名詞，用言をルールとして取り入れる．否定的な名詞，用言の辞書として，日本語評価極性辞書（用言編 [小林 05]，名詞編 [東山 08]）を用いる．加えて，否定的な名詞，用言のみで取得できない不満を抽出するために,「〜過ぎ」「〜にくい」「〜づらい」「〜ない」といったパターンを別途利用する．

本研究では，以上の手続きで整備した要望抽出ルールが自治体管理名詞句と同一ツイート内で共起した場合に，そのツイートを要望として抽出した．

5. 実験

5.1 実験設定

評価用データの作成には，2013 年 4 月 1 日からの 1 年間に投稿されたツイートを利用した．これらのツイートから自治体名（20 件の政令指定都市名）を本文中に含むものをランダムに 4,995 件取得し，加えて各々のツイートの前後 30 分間以内に投稿されたツイートを収集した．この際，BOT により自動的に投稿されたと思しきツイートの除去を事前に行っている．

このようにして収集された合計 41,978 件のツイートに対して，4.2 節の定義に従って，人手により「直接要求」「要求意図」「不満」のラベルを付与した．その結果,「直接要求」は 50 件,「要求意図」は 50 件，また「不満」は 350 件付与された．実験は，以下の 3 通りの設定で行った．

手法 (i) 自治体名の含まれるツイートのみを手法適用対象とする
自治体名の含まれるツイートのみに「直接要求」「要求意図」「不満」のラベルを付与し，それ以外のツイートには「その他」ラベルを付与する．

手法 (ii) 前後 30 分間以内のツイートも解析の対象にする（全てのツイートを対象）
自治体名の含まれるツイートと，その前後 30 分間以内の同一ユーザによるツイートの全てに「直接要求」「要求意図」「不満」のラベルを付与する．手法 (i)，(ii) を比較することで周辺ツイートを考慮することの有用性を確かめる．

手法 (iii) 前後 30 分間以内のツイートも解析の対象にする（自治体管理名詞句を含むツイートのみ対象）
自治体名の含まれるツイートと，その前後 30 分間以内の同一ユーザによるツイートのうち自治体管理名詞句を含むものに「直接要求」「要求意図」「不満」のラベルを付与する．手法 (i)，(iii) を比較することで周辺ツイートを考慮することの有用性を確かめ，手法 (ii), (iii) を比較することで自治体名の含まれるツイートの周辺ツイートを考慮する際の，自治体管理名詞句の有無による影響を調べる．

5.2 実験結果・考察

実験結果を表 2 に示す．これより，以下のことがわかる．

- 前後 30 分間以内のツイートも対象とした場合，自治体名の含まれるツイートのみを対象とした場合に比べて，各ラベルについての再現率が向上している．これより，自治体名を含むツイートの周辺ツイートまで含めて見ることにより，自治体に対するより多くの要望が抽出されることが確認され，周辺ツイートを考慮するという本手法の有効性が示された．
- 手法 (ii) と (iii) の結果を比較すると，自治体管理名詞句の含まれないツイートに対しても 4.3 節で作成したルールにより要望抽出を試みると，適合率が極端に落ちることがわかる．これは，自治体管理名詞句の含まれないツイートを対象とすることによって，自治体に全く関連しない要望を誤って抽出しているためである．このため，多様な事物に関する投稿がなされる Twitter を対象に要望抽出を行うにあたって，本研究で定義した自治体管理名詞句が有用であることがわかる．

各自治体に対して抽出された要望の例を表 3 に示す．このうち，横浜市に対する直接要求，京都市に対する要求意図，名古屋市に対する不満のそれぞれの例には自治体名が含まれないが，各々「バリアフリー」「市電」「バス」という自治体管理名詞句を含むため，抽出することができた．

5.3 エラー分析

本手法により抽出できなかった要望ツイートの例を以下に示す．

(10) [直接要求] メロディ導入するんならアナウンスに被らないようにその前に入れろよ

表 2: 手法 (i), (ii), (iii) による直接要求，要求意図，不満分類の評価結果

	直接要求			要求意図			不満			合計		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
手法 (i)	8.12 (22/271)	44.00 (22/50)	13.71	18.60 (16/86)	32.00 (16/50)	23.53	7.60 (151/1986)	43.14 (151/350)	12.93	8.07 (189/2343)	42.00 (189/450)	13.53
手法 (ii)	2.22 (40/1800)	80.00 (40/50)	4.32	5.90 (33/559)	66.00 (33/50)	10.84	2.08 (273/13156)	78.00 (273/350)	4.04	2.23 (346/15515)	76.89 (346/450)	4.33
手法 (iii)	8.96 (36/402)	72.00 (36/50)	15.93	17.57 (26/148)	52.00 (26/50)	26.26	6.94 (207/2984)	59.14 (207/350)	12.42	7.61 (269/3534)	59.78 (269/450)	13.50

表 3: 各自治体に対して抽出された要望の例

要望の種類	対象の自治体	抽出された要望の例
直接要求	札幌市	札幌市よ、ウチの周りのような住宅地の除雪が最悪なのだがなんとかせえや
	仙台市	職場周辺の悪臭が、去年の夏からひどい。仙台市さんよ、どうにかしてくれ。
	横浜市	帰宅困難者一時滞在施設検索システムですが、アイコンの図案が全て同じため色覚障害者には受入可と不可の区別が出来ません。バリアフリーの観点から図案変更の検討をお願いします
要求意図	大阪市	大阪市は南港の交通の便をお台場レベルにまで早急に引き上げるべき。
	広島市	広島市の東区役所って空いてるスペース売ればいいのに。 職員少ないんだからムダなスペース多い。
	京都市	市電をネタに金を稼ぎたいのなら、ターゲットをマニアに絞って、マニアから効率よく金を巻き上げることを追求した方がいいんじゃないだろうか。
不満	堺市	我が町、堺市の住民カードがダサすぎる件
	川崎市	今日も暖房が弱すぎる　川崎市バス
	名古屋市	バス内がうるさいのにはもう慣れた

(11) [要求意図] 東区役所を縮小してなにか新しくやれば
(12) [不満]30 分も帰る時間遅くなる

(10) の直接要求と (12) の不満は，それぞれ自治体管理名詞句を含まないツイートであったため，本手法で抽出できていなかった．また，(11) の要求意図は,「やれば」という要求意図表現が 4.3 節で定義したルールに含まれなかったため，抽出の対象外となっていた．

次に，本手法により誤って抽出したツイートの例を以下に示す．

(13) 横浜市水道局がアクセルワールドとコラボしたんやから、防衛省はガルパンとコラボすべき (適当)
(14) マリノスとベイスターズが全勝とか、全横浜市民が感動の涙に溺れるべき。

(13) と (14) は，各々4.3 節で作成された「すべき」,「るべき」という要求意図表現を含んでいるため，要求意図ツイートとして抽出された．しかしながら，(13) は「横浜市」に対してではなく「防衛省」に対する要求意図であり，自治体に対する要望として抽出するのは不適切である．また，(14) は対象が「全横浜市民」であるが，ツイートの内容を見ると一種の冗談であると考えられ，抽出するのは不適切だと考えられる．

6. おわりに

本研究では，Twitter を用いた自治体への要望抽出手法を提案した．その際に，時間的に近接するツイートも含めて見たうえで，自治体管理名詞句の辞書を利用し，ルールベースの要望抽出を行った．

今後の課題として，ツイートが自治体に関するものか否かの，より精緻な判別が挙げられる．本研究では自治体名あるいは自治体管理名詞句を含むツイートのみを要望抽出対象としていたが，5.3 節でのエラー分析の結果，本手法では取得できない要望も存在した．本手法ではツイート本文のみを要望抽出の手がかりとしていたが，プロフィール情報，位置情報付きツイートなどを考慮し，ユーザの居住地を推定する手法などを取り入れ，より多くの要望を抽出するべく検討したい．

謝辞

本研究は，東北大学工学部　情報知能システム総合学科「Step-QI スクール」の支援を受けた．

参考文献

[Boettcher 12] Boettcher, A. and Lee, D.: EventRadar: A real-time local event detection scheme using twitter stream, in *Green Computing and Communications (GreenCom)*, pp. 358–367 (2012)

[永野 12] 永野峻祐, 小根山裕之, 大口敬, 鹿田成則：形態素解析を用いたアンケート調査自由記述欄の分析手法に関する研究ー路面電車利用意識調査データを用いたケーススタディー, 土木学会論文集D 3（土木計画学）, Vol. 68, No. 5, pp. 973–981 (2012)

[山本 06] 山本瑞樹, 乾孝司, 高村大也, 丸元聡子, 大塚裕子, 奥村学：文章構造を考慮した自由回答意見からの要望抽出, 言語処理学会第 12 回年次大会 (2006)

[山本 12] 山本修平, 佐藤哲司：Twitter からの実生活情報の抽出法の提案, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2012) F3-4 (2012)

[小林 05] 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. 2, pp. 203–222 (2005)

[大塚 04] 大塚裕子, 内山将夫, 井佐原均：自由回答アンケートにおける要求意図判定基準, 自然言語処理, Vol. 11, No. 2, pp. 21–66 (2004)

[渡辺 11] 渡辺一史, 大知正直, 岡部誠, 尾内理紀夫：Twitter を用いた実世界ローカルイベント検出, 第 4 回楽天研究開発シンポジウム予稿集 (2011)

[土屋 13] 土屋圭, 豊田正史, 喜連川優：マイクロブログを用いた鉄道の運行トラブル状況抽出に関する一検討, 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2013, No. 31, pp. 1–6 (2013)

[東山 08] 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)