過去の投稿を活用したマイクロブログユーザの現在位置推定

鈴木　有†　鍛治　伸裕††,†††　吉永　直樹††,†††　豊田　正史††

† 東京大学 大学院情報理工学系研究科　〒 113-8654 東京都文京区本郷 7-3-1
†† 東京大学 生産技術研究所　〒 153-8505 東京都目黒区駒場 4-6-1
††† 独立行政法人 情報通信研究機構　〒 184-8795 東京都小金井市貫井北町 4-2-1
E-mail: †{ysuzuki,kaji,ynaga,toyoda}@tkl.iis.u-tokyo.ac.jp

あらまし　マイクロブログの投稿には，実世界に関するリアルタイム性の高い情報が含まれているが，イベントや災害など，空間的局所性の高い情報を利用するためには，その情報がどの場所にいるユーザから発信されたか知っておくことが重要となる．本研究では従来手法に倣って位置推定タスクを多クラス分類問題として定式化し，位置推定対象の投稿に含まれる名詞の Bag of Words (BOW) に加えて，ユーザの過去の投稿から位置推定のための素性を抽出する．具体的には，過去の投稿中の名詞の BOW，及びその投稿と推定対象の投稿との時間近接性や名詞 BOW と係り受け関係にある動詞，時間表現を素性に含めることで，より高精度なマイクロブログユーザの位置推定を行う手法を提案する．実験では geotag 付きツイートに提案手法を適用し，現在の投稿の BOW のみを素性に用いるベースラインと比較して，平均誤差距離が都道府県レベルで 45.4km，市区レベルで 38.1km 減少し，提案手法の有効性を確認できた．

キーワード　geolocation, 位置情報, マイクロブログ,twitter

1. はじめに

近年，twitter に代表されるマイクロブログや foursquare など，ユーザの位置情報が紐付いたデータが拡大している．このような位置情報を有効活用することができれば，ユーザに適応した情報提供を行うことが可能になる．例えば，レストランやイベント等を推薦する際にはユーザの嗜好や属性などに加えて，ユーザの現在の位置情報を考慮することが重要となる．一方で，発言時のユーザの位置情報が利用可能になれば，それらのユーザの発信する情報を集約することで人々の動きを定量的に把握することができるようになり，道路や公共施設の建設計画のために活用できると期待される．さらに，SNS 等の投稿から取得できるユーザの趣味嗜好や属性情報なども合わせて利用することで観光施策や商業施設の出店計画，マーケティング等に活用することも可能となると考えられる．

このように，位置情報は様々な事象の予測やマーケティングに使える有益な情報であるが，自らの位置情報を積極的に公開しているユーザは少ない．そのため，マイクロブログテキストの内容から，ユーザの現在位置を推定するための技術に関する研究が盛んに行われている [1] [2].

しかしながら，マイクロブログテキストの多くは短く，ユーザの現在位置を推定するための情報が，必ずしも十分には含まれていない．例えば，図 1 下のツイートは新宿にいるユーザが発信したものであるが，このツイート単体から，ユーザの現在位置を推定することは困難である．ところが，この場合でも前日に図 1 上のようなツイートが発信されていたとすると，ユーザが現在，新宿に滞在していることを推測することができる．本研究では，このようにユーザの過去の投稿を適切に利用することによって，ユーザの位置推定の精度を高める方法について検討を行う．

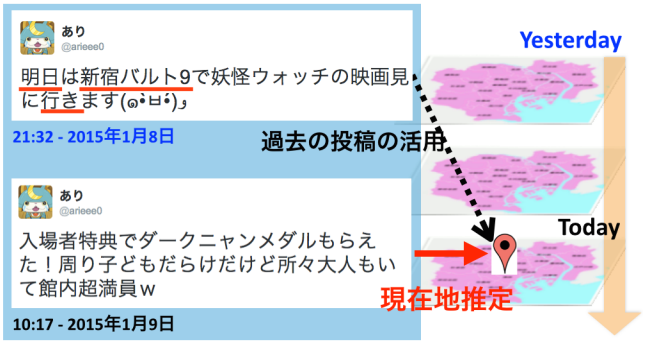

図 1　ユーザ現在位置推定における過去の投稿の活用

過去の投稿にはユーザの現在位置を推定するために有用な情報も含まれているが，現在の位置と無関係なノイズの情報も含まれている．このようにノイズを含んだ過去の投稿を現在位置推定に効果的に活用するため，本研究では「時間近接性」と「予定表現」という 2 つの観点に着目した新たな素性を提案する．

時間近接性とは推定対象の投稿に対して過去の投稿が時間的にどれほど近接しているかの時間差を表し，2 つの点で現在位置推定に活用することができる．1 つは時間差が小さければ小さいほどユーザの現在位置と過去の投稿内容との関連性が高くなるという移動の物理的制約を情報として活用できる点である，もう 1 つは Bag of Words (BOW) 素性ごとに異なる滞在時間や人の流動等の場所の特徴を部分的に活用できる点である．例えば，ラーメン屋など飲食店と比較すると，温泉レジャー施設などの滞在型スポットにおける典型的な滞在時間は長くなることが考えられる．また，浅草寺を訪れた人は，数十分から数時間後には東京スカイツリーを訪れる傾向が高いなど，典型的な行動パターンが存在すると考えられる．推定対象の投稿と時間的に近接した過去の投稿から素性を抽出することによって，こ

うした情報を学習することができると期待できる．

またユーザが未来の行き先について述べるような予定表現を現在位置推定に活用するため，「BOW と係り受け関係にある動詞」と「時間表現」を新たに素性として追加する．「係り受け動詞」素性は BOW に対してユーザが時空間的にどのような関係性にあるかを知る手がかりになる．例えば「渋谷」という BOW 素性に対して，格助詞「に」を介して動詞「いる」が係り受け関係にあるとすれば，ユーザが特定の位置 (例:渋谷区) に現在いることが示唆されるが，動詞「行く」と係り受け関係にあるとすれば，今は「渋谷」にいないが一定の時間後に「渋谷」にいることが示唆される．「時間表現」素性はユーザがある場所に存在する時間について知る手がかりになる．例えば「明日渋谷で買い物したい」という文では，名詞「渋谷」と時間表現「明日」の組み合わせ素性を使うことによって，ユーザが特定の位置（例:渋谷区）に存在する可能性が高い時間帯（例:翌日 0 時から翌日 24 時まで）について学習することが期待できる．

実験では，twitter ユーザが投稿した位置情報 (geotag) 付きツイートをデータセットとして用いて，ユーザの現在位置の推定を行った．位置推定はユーザが存在する行政区分を予測することで行い，行政区分は都道府県・市区の 2 種類を用いた．推定対象のツイートに含まれる BOW のみを素性とするベースラインと比較して，都道府県・市区レベルともに平均誤差距離が減少し，提案手法の有効性を確認できた．

本論文は，以下の構成に従う．まず 2. 節でユーザの位置推定について関連する論文について述べる．3. 節では，過去の投稿を考慮した提案手法について述べ，4. 節で評価実験とその考察を行う．以上を 5. 節でまとめる．

2. 関 連 研 究

本節では，マイクロブログの投稿からのユーザの位置を推定する既存手法について説明する．

ユーザの位置をユーザの投稿したテキストから推定する最も簡単な方法は，テキスト中に含まれる地名・場所名をあらかじめ用意した地名辞典などとパターンマッチさせることで場所を推定するものである．しかし，この方法では新たな市町村や施設などが誕生するたびに地名辞典を更新する必要がある上に，略語や特定の場所に関連が深い語（方言や，特産物，地域で話題のトピックなど）に対応できない．

そこで，位置情報が付随されているユーザのテキストを学習データとして用いることで，最新の地名や地域トピックなどもユーザの場所推定に利用できるようにする機械学習アプローチがよく用いられている．具体的には，ユーザの位置推定タスクを，ユーザの投稿をあらかじめ定めた地域区分のラベルの 1 つに分類する多クラス分類問題として捉え，一部の投稿に付加された geotag を教師データとして多クラス分類器の学習を行う．位置推定に用いる素性としては，位置推定対象の投稿に含まれる単語，すなわち BOW を用いる研究が多い [3] [4]．

これらの研究では，位置推定対象の投稿のみから素性の抽出を行っているが，マイクロブログでは一つの投稿の文字数が制限されている場合が多く，位置推定に必要となる情報を十分に含んでいないことが多々ある．結果として，市町村レベルの位置推定精度は頭打ちになっているのが現状である [1]．

この問題に対し，伊川ら [2] は，ある投稿から遡って 10 分以内の投稿であればほぼ同じ場所であると考え，過去 10 分以内に foursquare による投稿があればそれらを学習に利用している．また，Hong [5] らは，ユーザの位置情報付き投稿からユーザの出現位置の分布をモデル化し，これを事前分布として生成モデルに組み込むことで投稿時のユーザの位置推定に活用している．

本研究では，位置情報付き投稿に限らず一定期間内の全ての過去の投稿を利用する．その際，過去の投稿と位置推定対象の投稿時間差を考慮に入れることで，過去の投稿からより多くの特徴量を拾い上げることを試みる．さらにこれに加えて，「明日○○に行く」「土日に××に行きたい」など文中に含まれる時間表現や，BOW と係り受けの関係がある動詞・モダリティを考慮し，素性に含めることで，過去の投稿を有効にユーザの現在位置推定に活用することを試みる．

3. 提 案 手 法

本節では，過去の投稿を利用したマイクロブログユーザの位置推定手法を提案する．

あるユーザの過去の投稿から抽出した BOW 素性は，そのユーザの現在位置推定にも有効であるというのが提案手法の基本的な考え方であるが，実際にはユーザの現在位置と無関係な投稿も多く，単純に BOW 素性として追加するだけでは誤分類を招くことが予想される．そこで本研究では，「時間的近接性」と「予定表現」を考慮することによって，過去投稿の情報のより高度な活用を実現する．時間近接性とは位置推定対象の投稿と過去の投稿がどれだけ時間的に近いかを示す時間差であり，過去投稿から抽出した BOW 素性と時間差を組み合わせた素性を用いることで，移動の物理的制約などを学習することが期待できる．また，予定表現とは未来の行動予定に関する言及であり，そうした具体的な予定表現に着目することによって，推定対象の投稿との時間差が大きくても，現在位置推定に有効活用することが可能になると期待できる．本研究では具体的に予定表現をとらえる要素として「BOW と係り受け関係にある動詞」と「時間表現の有無」に着目した．これらの手がかりを，多クラス分類器の素性として表現し，従来の BOW 素性に追加する．

なお，多クラス分類器としてはサポートクラス Passive Aggressive アルゴリズム I (SPA-I) [6] を用い，実装には opal(注1) を用いた．また，BOW 素性としては，投稿を MeCab (注2)(IPA 辞書) で形態素解析して得られた結果から，名詞句を取り出して利用した．名詞句は J.DepP (注3) によって同文節と判定された名詞の中で連続したものを結合し，作成する．

3.1 時間近接性に着目した過去の投稿の利用

位置推定対象の投稿に対し，過去の投稿との時間差が小さけ

(注1)：http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/
(注2)：https://code.google.com/p/mecab/
(注3)：http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

れば小さいほど，移動の物理的制約から投稿時のユーザの位置も近いと考えられる．このような時間近接性とユーザの位置の関連度を捉えるため，ユーザの過去の投稿を複数の時間窓に重複を許して分類し，それぞれの投稿の集合から BOW 素性を抽出する．具体的には，過去の投稿を，5 分以内，20 分以内，75 分以内，5 時間以内，21 時間以内，80 時間以内からなる 6 の時間窓に分類した．これら過去の投稿から抽出された BOW 素性とそれらが発信された時間窓を合わせたものを，以下では時間情報付き BOW 素性と呼ぶ．

このように過去の投稿との時間差と BOW を組み合わせることで単語ごとに異なる，投稿時間差とユーザの現在位置への関係を考慮できることに注目されたい．例えば，遊園地（例: ディズニーランド）のような滞在時間が長い施設に関する BOW 素性は，時間差「21 時間以内」との組み合わせに対して，特定の位置（千葉県浦安市）との相関が期待できるし，一方でレストランのような滞在時間が短い施設に関する BOW 素性については,「75 分以内」などより短い時間差との組み合わせについて相関が高くなると期待できる．

3.2 予定表現に着目した過去の投稿の利用

ユーザは必ずしも現在位置に関連した投稿をするわけではなく，過去や未来の行き先について述べている投稿もある．埼玉県の秩父市に言及しているツイートを対象とした予備調査を通して，過去や未来の行き先について述べている投稿が多く存在していることを示した上で，それらを情報として活用するために提案する，予定表現に着目した素性について述べる．

筆者らはツイート本文に「秩父」を含むツイートをランダムに抽出 (注4) し，ユーザらがどのような意味で「秩父」について言及しているかを表 1 のように 8 つのカテゴリにラベル付けし，そのツイート数を調査した．結果は表 2 に示す．

表 2 を見ると,「秩父」を含むツイートの中に未来の予定に関するツイートが 145 件もある．これは現在の 189 件と比べても少なくなく，予定表現を述べられた投稿はある程度の数存在し，ユーザの現在位置推定をする上での情報として活用できると思われる．

予定に関して述べている過去の投稿をユーザの現在位置推定に活用するため，BOW と係り受け関係にある動詞と，投稿に含まれる時間表現を素性として加える．

3.2.1 BOW と係り受け関係にある動詞の利用

BOW と係り受け関係にある動詞を素性に加えることで投稿中の BOW とユーザの位置関係を学習する．

前節で行った調査実験の結果ではユーザの現在位置としてだけではなく，過去や未来の行き先などに言及する投稿も数多くあることが分かった．そのような投稿に対して，単純に文中に含まれる BOW を素性とした場合，誤分類につながる．そこで，BOW と係り受け関係にある動詞を素性に加える．例えば,「秩父（に)」の係り先が「帰りたい」や,「行ってきた」であれば，そのユーザは秩父にいないはずである．一方で，係り先が「いる」や「来た」であれば秩父にいることが期待できるので，上記の素性により，これらの差異を捉えられるようになると期待できる．

素性がスパースになることを避けるため，素性に含める動詞は表 3 に示すような，移動を表す動詞，場所と共起しやすい動詞に限定した．また素性には願望（〜したい），過去（〜した），否定（〜しない）の 3 種類のモダリティや，名詞句の直後に助詞（に・で・へ・にて・まで・から・を）が存在する場合は係り受け動詞に付随させた．

表 1　秩父に関する投稿の分類

分類種類	基準
現在	現在秩父にいることを示すツイート．そのユーザがツイートをした時刻，または直近 (数分以内) にいると思われるもの
未来	将来訪れることがほぼ確実，具体的な予定あり
過去	過去秩父を訪れたことを示すツイート．回想ではなく，訪れた具体的な時間が推定できるもの
願望	具体的な予定なし．ぼんやりとした希望，迷い
話題 (回想含む)	自分がいる場所・行く場所とは関係無いもの．回想はこれに含む．
情報・宣伝	自分がいる場所・行く場所とは関係無いもののうち，ユーザの感想を含まず，何らかの告知や情報を提供する目的のもの
秩父に関係なし	秩父（秩父地方・秩父郡含む）に関係ないもの
不明	上記に分類できないもの

表 2　秩父に関する投稿のタグ付け結果

分類種類	数	割合	ツイート例
現在	189	6.3%	おはようございます！あの花舞台の秩父でキャンプしております
未来	145	4.8%	今から秩父行ってくる (送別会の為)
過去	79	2.6%	今日？今日は秩父にあそびにいったよ★似てる人いた？
願望	169	5.6%	夏休みに秩父に行く人多そうな気がする…というか私が行きたい
話題 (回想含む)	1297	43.2%	バナナマンしたら秩父出身 wwwwwwww
情報・宣伝	527	17.6%	埼玉県の天気→秩父地方 (秩父)8 日 (土) 曇のち晴 27 ℃/15 ℃
秩父に関係なし	526	17.5%	秩父宮ラグビー場って東京？
不明	68	2.3%	秩父と
合計	3000	100%	

(注4)：なおデータの取得期間は 2011 年 4 月から 2013 年 3 月の 2 年間で,「秩父」を含むという条件は 343,054 件のツイートが該当した．この中から 3000 件をランダムに選んでいる．

表 3　素性に含める動詞

来る, くる, 行く, いく, 着く, つく, 歩く, 出る, でる, 過ぎる, すぎる,
戻る, もどる, 帰る, 降る, 降りる, おりる, 入る, 向かう, 止まる,
迷う, 急ぐ, 移動する, 停る, 通る, いる, 居る, ある, なる, する, やる,
飲む, 乗る, 行う, 食べる, 食う, 会う, 混む, わかる, 打ち合う,
出来る, 開催する, 寄る, 呑む, 見る, 抜ける, いたす, あふれる, 探す,
溢れる, 落ち着く, 潰す, 変わる, 盛り上がる, 減る, みる, 見つける,
配布する, 始まる, おる, 売る, 観る, できる, 繋がる, 貰う, 集める,
待つ, 配る, 買い物する, 賑わう, 上がる, オープンする, 始まる,
見る, 占領する, 見える, 見かける, 頂く, 似合う

表 4　定量的時間表現

時間表現	一致単位
今	現在
現在	現在
なう	現在
明日	日
あした	日
明後日	日
あさって	日
土日	週
今週	週
来週	週

表 5　感覚的時間表現

これから, もうすぐ, いまから, 今から, いまさら, いまだに, かつて,
急に, きゅうに, 最近, さきほど, さっき, 早速, さっそく, しばらく,
すかさず, すぐに, そのうち, ただいま, ただちに, だんだん, 近頃,
ちかごろ, とうとう, とりあえず, のちほど, 後ほど, あとで, 後で,
もう少ししたら, もうちょっとしたら, あと少ししたら, ほどなく,
まもなく, もう, もうじき, ようやく

3.2.2　投稿に含まれる時間表現の利用

文中の BOW と共起する時間表現を素性に含めることで，ユーザの予定情報のうち，ユーザがある場所に存在する時間帯に関する情報を抽出する．時間表現には「定量的時間表現」と「感覚的時間表現」の 2 種類があるとして，それぞれ異なった方法で素性を作成する．

定量的時間表現とは「明日」「来週」など時間表現が指す時間範囲を定量的に表すことができる時間表現である．ユーザの過去・現在の投稿に定量的時間表現が含まれていた場合は，時間表現が示す時間範囲が推定対象の投稿が発信された時間を含むかどうかの真偽値と，その一致単位 (現在, 日, 週) を素性とした．本研究で使用した定量的時間表現を表 4 に示す．

感覚的時間表現とは「もうすぐ」「これから」などそれらが指す時間範囲が筆者の主観で変わりうる時間表現である．ユーザの過去・現在の投稿に感覚的時間表現が含まれていた場合は，その時間表現そのものを素性に含めた．前述した時間窓と組み合わさることで，たとえば「もうすぐ」という感覚的時間表現は「21 時間以内」よりも「20 分以内」の時間窓と強い関連性がある，といったように，感覚的時間表現が一般的に指す時間範囲をうまくとらえることができると期待する．本研究で使用する感覚的時間表現を表 5 に示す．

表 6　データセット

	訓練	開発	テスト
ユーザ数	34,400	4,306	4,306
ツイート数	5,554,630	680,997	692,005

4. 評価実験

我々の研究室において 2012 年 7 月から 2014 年 6 月の期間に収集した geotag 付き日本語ツイートから，次のようにして実験データセットを構築した．Han [7] らの研究に従い，上記期間において geotag 付きツイートを 10 回以上投稿したユーザを抽出した．次に，その中から，bot 等のユーザを排除するため，投稿回数が 10,000 回以上であるユーザを削除すると，最終的に 43,000 人のユーザと 9,887,995 ツイートが得られた．

geotag 付きツイートには位置情報サービスを通して作成されたツイートが多く，これらは決まった投稿文の形式で地名等が書かれているためユーザの位置推定を過度に容易にしてしまうため，データセットからは削除する．削除した位置情報サービスは foursquare (注5)，ロケタッチ (注6)，ついっぷるスポット (注7)，今ココなう！(注8) の 4 つである．位置情報サービスによるツイートは foursqure の「I'm at ..」など決まりきった投稿文の形式にパターンマッチするか否かで判定した．位置情報サービスによるツイートを除去した残りの 6,927,632 ツイートをデータセットとして使用する．

データセットはユーザ数ベースで 8:1:1 の割合で分割し，それらのユーザによるツイートをそれぞれ訓練データ，開発データ，テストデータとした．データセットの訓練データ，開発データ，テストデータにおけるユーザ数とツイート数を表 6 に示す．

なお，opal の学習時の繰り返し回数のパラメータは 5 回とし，カーネルは線形カーネルとした．オプションは学習データのシャッフリング (-s) とパラメータの平均化 (-a) を設定し，開発データを用いて -c オプションによるチューニングを行った．

分類のラベルに用いた地域区分は，国土交通省による平成 25 年度大字・町丁目位置参照情報 (注9) を参考に，日本の 47 都道府県および日本の全 1903 市町村のうち 962 市区を利用した．東京 23 区や政令指定都市などは 1 つの市ではなく，各区をそれぞれ 1 つの行政区域として扱っている．全 1903 市町村から 941 町村を含めず 962 市区だけ利用した理由としては，twitter による投稿が市区などの都市圏に集中していることがあげられる．用いるデータセットの訓練データで調査したところ，962 市区からのツイートが 94.7%を占めていた．そのため，クラス数に比例したメモリが必要となる SPA-I アルゴリズム による学習を可能とするため，市町村レベルのラベルは市区だけを用いることにした．なお訓練データからは 962 市区以外の 941 町

(注5)：https://ja.foursquare.com/
(注6)：http://tou.ch/
(注7)：http://s.twipple.jp/
(注8)：http://imakoko-gps.appspot.com/
(注9)：http://nlftp.mlit.go.jp/isj/

村に割り当てられたツイートは削除したが，開発データ，テストデータにおいて 941 町村に割り当てられたツイートはそのまま残し，全て誤分類したものとして扱った．

ツイートの geotag から得られる緯度経度情報を行政区分へと変換するには，上述の各市区の大字・町丁目の緯度経度情報の中で最も近いものに割り当てるという方法を用いた．訓練データからランダムに取得した 1000 ツイートについて，Yahoo!リバースジオコーダ API (注10) による市町村割り当て結果を正解として，上記の方法の精度を検証すると市町村単位で 99.0%，都道府県単位で 99.9%であった．また，誤った市町村名を割り当てられたツイートを人手で調べたところ，全て正しい市町村から数十 m 程度離れた場所で投稿されていた．このことから，十分な精度で市区名が割り当てられていると考えられる．

4.1 評価指標

評価指標は行政区分の予測精度に関する指標と実際のユーザ位置との誤差距離に関する指標の 2 タイプ用意した．これはアプリケーションによってユーザの行政区分が知りたい場合や，緯度経度などユーザが存在する詳細な位置を知りたい場合など，知りたいユーザ位置の種類が異なることが考えられるためである．

行政区分の予測精度に関する指標では，正解ラベルと予測ラベルが一致した精度（Acc）と，正解ラベルが予測ラベルの N km 以内に存在した場合の精度 （Acc@Nkm）を用いる．市区レベルの推定の場合，962 クラスの分類問題にあたるため予測ラベルと正解ラベルを完全に一致させるのが難しい．そのため Acc@Nkm を指標として用いる．

ユーザ位置との誤差距離に関する指標は，予測ラベルの行政区分の代表点の緯度経度と，テストデータの投稿に付加されているユーザ位置の誤差の平均値と中央値を用いる．予測ラベルの行政区分の代表点としては平成 24 年の総務省における人口重心データ (注11) を用いた．

4.2 実験結果

ユーザ位置推定について，都道府県レベルの結果を表 7 に，市区レベルの結果を表 8 に示す．過去 BOW，係り受け，時間表現はそれぞれ提案した 3 つの素性である，時間情報付き BOW(3.1 節)，BOW と係り受け関係にある動詞 (3.2.1 節)，時間表現 (3.2.2 節) をベースライン素性に追加したことを意味する．ベースラインは，学習データで最も頻出した地域区分にテストデータを全て割り当てる最頻地域区分と，過去の投稿を一切使わずに，位置推定対象の投稿から抽出した BOW 素性のみを考慮する現在投稿 BOW を用いた．

結果は，提案手法のうち，時間情報付き BOW（過去 BOW）と係り受け関係にある動詞をベースライン素性に追加した場合に最も位置推定精度が高くなり，都道府県レベルのユーザ位置推定ではすべての指標で，市区レベルのユーザ位置推定では Acc 以外の指標でベースラインよりも位置推定精度が改善した．

(注10)：http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/reversegeocoder.html

(注11)：http://www.stat.go.jp/data/kokusei/topics/topi61.htm

表 7　実験結果 (都道府県レベル)

素性	分類精度		誤差距離	
	Acc	Acc@50km	平均値	中央値
最頻地域区分	28.71	52.1	262.7	241.0
現在投稿 BOW	43.00	52.61	218.13	45.78
＋ 過去 BOW	51.85	62.23	173.80	25.20
＋ 過去 BOW ＋ 係り受け	**52.02**	**62.40**	**172.77**	**24.92**
＋ 過去 BOW ＋ 時間表現	51.82	62.19	173.88	25.16
＋ 全て	51.98	62.37	172.95	24.86

表 8　実験結果 (市区レベル)

素性	分類精度		誤差距離	
	Acc	Acc@50km	平均値	中央値
最頻地域区分	12.78	23.55	301.84	357.79
現在投稿 BOW	**23.63**	40.10	267.28	170.37
＋ 過去 BOW	21.24	48.82	232.65	58.22
＋ 過去 BOW ＋ 係り受け	21.72	**49.61**	**229.14**	**52.51**
＋ 過去 BOW ＋ 時間表現	20.42	48.10	236.13	64.16
＋ 全て	20.25	48.25	235.95	62.91

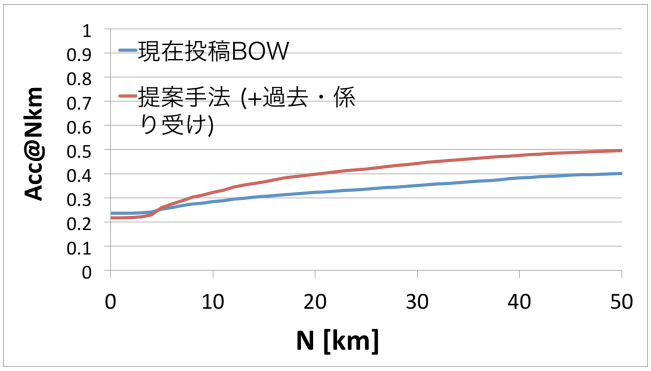

図 2　市区レベルにおける Acc@Nkm

特に誤差距離では現在投稿 BOW に比べて提案手法では都道府県レベルで平均値が 45.4km，中央値が 20.9km 減少し，市区レベルで平均値が 38.1km，中央値が 117.9km 減少した．

提案手法の地域区分の予測傾向についてより詳細に調べたものとして，市区レベルの Acc@Nkm についてのグラフを表 2 に示す．正解ラベルの Nkm 以内なら正解とする Acc@Nkm において，N=5 以上において，提案手法がベースラインを上回っている．この要因として提案手法は過去の投稿を利用するため使用できる情報量が多く，誤分類する場合でも周辺の地域区分を予測しやすい傾向にあることが考えられる．

また，ベースラインの現在投稿 BOW と提案手法を用いたユーザ現在位置予測の結果について，都道府県の正解ラベルの分類先を日本地図にマッピングしたものを図 3，4 に示す．ここでは福島県が正解ラベルである例を掲載した．色の濃淡は分類数の多さによって 5 段階に分けられて塗られており，濃い色が分類数が多いことを示す．正解ラベルが福島県のデータに対する分類先について，ベースラインに比べて提案手法では東日本，特に東北地方や北関東などに位置する周辺の都道府県に分

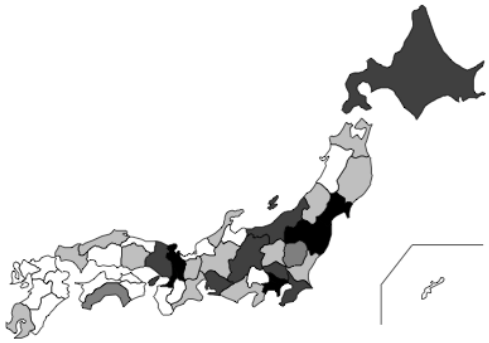

図 3　正解ラベルが福島県のテストデータの分類先 (素性:現在投稿 BOW)

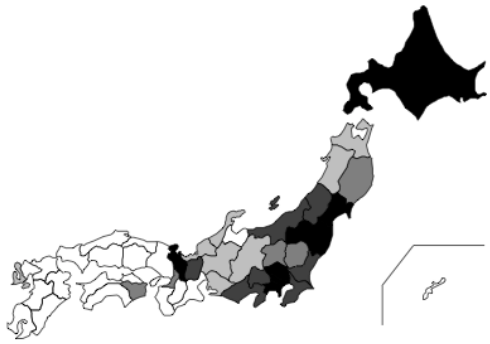

図 4　正解ラベルが福島県のテストデータの分類先 (素性:提案手法)

類されやすくなっていることがわかる．この例からも提案手法では，過去の投稿を考慮して現在位置を推定するため使える情報が多く，ベースラインに比べて誤分類時でも周辺の都道府県を予測しやすい傾向にあることが確認できる．

4.3 考　　察

提案した素性のうち，時間情報付き BOW 素性を加えた場合に誤差中央値距離や Acc@50km などの指標で大幅に精度が向上しており，特に有効な素性であると考えられる．

時間情報付き BOW 素性が有効に働いた要因として，時間近接性を考慮することができたことや，BOW 素性ごとに滞在時間等の場所の特徴を部分的に考慮することができたことが考えられる．

時間近接性は推定対象の投稿と過去の投稿との時間差が小さければ，ユーザの現在位置と過去の投稿に含まれる BOW との関連性が高いという仮定であるが，時間情報付き素性はこれらの関係性を素性の重みとして反映させている．素性の設計上，時間情報付き素性は重複を許して発火するので，過去の投稿との時間差が小さいほどその投稿に含まれる BOW については多くの時間情報付き素性が作られる．例として「東京タワー」という BOW に関する時間情報付き素性について，訓練データの学習によって付けられた重みを表 9 に示す．例えば，推定対象の投稿の過去 50 分前に「東京タワー」を含む投稿があれば，東京タワー (75 分以内)，東京タワー (5 時間以内)，…，東京タワー (80 時間以内) の 4 つの時間情報付き BOW 素性が全て発火するが，過去 80 時間前にしか「東京タワー」を含む投稿がなければ東京タワー (80 時間以内) の 1 つの素性しか発火しない．よって過去 80 時間前に「東京タワー」を含む投稿がある場合より過去 50 分前にそのような投稿がある場合の方が東京都-港区に対する重みがはるかに大きくなり，投稿時間差とユーザの位置の関係を反映できることがわかる．

表 9 「東京タワー」に時間情報が付随した素性の重み

素性	東京都-港区の重み
東京タワー (現在)	0.960
東京タワー (75 分以内)	0.231
東京タワー (5 時間以内)	0.107
東京タワー (80 時間以内)	0.062

表 10　各 BOW に対する時間情報付き素性の重み

素性	所在地	重み
東京駅 (現在)	東京都-千代田区	0.533
東京駅 (20 分以内)		0.213
東京駅 (75 分以内)		0.192
東京駅 (5 時間以内)		0.055
秋葉原 (現在)	東京都-千代田区	0.749
秋葉原 (75 分以内)		0.255
秋葉原 (5 時間以内)		0.287
秋葉原 (21 時間以内)		0.064
ディズニーランド (現在)	千葉県-浦安市	0.343
ディズニーランド (5 時間以内)		0.137
ディズニーランド (21 時間以内)		0.174
ディズニーランド (80 時間以内)		0.066

また，時間情報付き BOW 素性を加えることで，BOW 素性ごとに滞在時間等の場所の特徴を部分的に学習できていることが精度向上に寄与していると思われる．例として「東京駅」「秋葉原」「ディズニーランド」に関する時間情報付き BOW 素性とそれらの所在地の市区に対して学習された重みを表 10 に示す．これら 3 つの施設・地名に対して最も重みが大きく付くような時間情報はそれぞれ異なる．東京駅は東京都千代田区に存在する日本を代表するターミナル駅であり，乗り換えで使われることが多いため，場所の性質として訪れる人の滞在時間はせいぜい数十分の滞在が予想される．それらを反映して，東京駅に関する時間情報付き素性は 75 分以内から 5 時間以内にかけて重みが急速に小さくなっており，東京駅を訪れる人は数十分程度の滞在が多いことが反映されている．一方，秋葉原やディズニーランドなど，東京駅よりも滞在時間が長いと思われる場所に関しては，75 分や 5 時間が経っても重みに大きな減少はみられなかった．

ベースライン（現在投稿 BOW）に時間情報付き BOW 素性のみを加えた場合に比べて，BOW と係り受け関係にある動詞をさらに素性に加えた場合では，精度が都道府県レベルで 0.17%，市区レベルで 0.48%向上した．係り受け動詞素性を加えることで，ユーザと投稿文中の BOW との位置の関係性を部分的に捉えられることが精度向上に寄与していると思われる．時間情報付き BOW に係り受け動詞を素性として加えたことにより位置推定が改善したツイート例（都道府県レベル）を表 11 に，関連する素性の重みを表 12 に示す．この例では，79 時間

表 11　係り受け動詞素性が効果的な例

時間区分	時間差	投稿内容抜粋
過去	79.1 時間前	夜行バスなう〜大阪に帰りまぁす バイバイ渋谷
現在	-	(MENTION) おはよ (略) 早起きやね
正解ラベル	大阪府	

表 12　表 11 に関する素性の重み

素性	重み最大ラベル	重み
渋谷 (80 時間以内)	東京都	0.318
大阪 (80 時間以内)	大阪府	0.154
大阪 (80 時間以内) +に+帰る	大阪府	0.180

表 13　時間表現素性が効果的な例

時間区分	時間差	投稿内容抜粋
過去	12.9 時間前	今日は事務所勤務〜そして、終わったら、渋谷の DUO へ
現在	-	fracoco の社長と Naa と
正解ラベル	東京都-渋谷区	

表 14　表 13 に関する素性の重み

素性	重み最大ラベル	重み
六本木 (80 時間以内)	東京都-港区	0.117
六本木 (80 時間以内) +に +行く	東京都-港区	0.000
渋谷 (21 時間以内)	東京都-渋谷区	0.075
渋谷 (80 時間以内)	東京都-渋谷区	0.094
渋谷 (21 時間以内) +予定一致 (日単位)	東京都-渋谷区	0.043
渋谷 (80 時間以内) +予定一致 (日単位)	東京都-渋谷区	0.090

前の過去の投稿に「大阪」「渋谷」という 2 つの地名表現が含まれているが，大阪に対してのみ「(に) 帰る」という動詞が係っており，大阪がこれからの移動の目的地であることを示唆している．係り受け動詞素性を加えることで,「渋谷 (80 時間以内)」「大阪 (80 時間以内)」という時間情報付き BOW 素性に加えて「大阪 (80 時間以内) +に+帰る」という，大阪という BOW に対する位置関係を考慮に入れた素性が発火し，予測ラベルとして大阪府に対する重みを大きくすることができる．

ベースライン（現在投稿 BOW）に時間情報付き BOW 素性のみを加えた場合に比べて，時間表現をさらに素性に加えた場合では，精度は向上しなかった．しかし時間表現素性を加えることで，ユーザがある場所に存在する確度の高い情報を部分的に抽出できている場合もみられた．時間情報付き BOW に時間表現を素性として加えたことにより位置推定が改善したツイート例（市区レベル）を表 13 に，関連する素性の重みを表 14 に示す．この例では，位置推定対象の投稿の 12 時間前に「渋谷」という地名が含まれている投稿があり，同文中に「今日」という予定表現が共起している．12 時間前の投稿における「今日」が指す一日は推定対象の投稿が発信された日に一致するので,「渋谷 (21 時間以内)」といった時間情報付き BOW 素性に加えて,「渋谷 (21 時間以内) +予定一致 (日単位)」といった時間表現を考慮した素性が発火し，予測ラベルとして東京都-渋谷区に対する重みを大きくすることができる．

表 15　人手によるツイート位置ラベル付け

	正解数	不正解数 誤り	不明
最頻地域区分	25	75	-
現在投稿 BOW	34	66	-
人手によるラベル付け	29	7	64

提案手法において予測できなかったツイートに最も多く見られた特徴として，ツイート内に位置に関する情報が少ないことがあげられた．位置情報が少ないツイートがどの程度存在するかを確認するために以下の調査実験を行った．まず開発データからランダムに 100 件ツイートを選び，そのツイートのみを見て人手で都道府県レベルでユーザ位置を推定した．結果を表 15 に示す．正解は 100 件中 29 件にとどまり，現在投稿 BOW ベースラインよりも少ない正答数となった．不正解となったツイートの内，ツイート内容を見ても位置が全く推定できない「不明」に 64 件が分類され，人が見ても場所を推定できないような情報量が少ないツイートが数多く存在することがわかる．このことから今後は地名などを含むツイートのみを位置推定対象とすることや，ツイートごとに位置推定がどの程度確からしいかの確度を導出することなどが必要であると思われる．

5. まとめと今後の課題

本研究では,「時間近接性」と「予定表現」に着目することでユーザの過去の投稿を現在位置推定へと活用した．ベースラインの BOW 素性による分類と比較して,「時間情報付き BOW」「係り受け関係にある動詞」の 2 つの素性を加えると平均誤差距離が都道府県レベルで 45.4km，市区レベルで 38.1km 減少し，提案手法の有効性を確認できた．

時間情報付き BOW 素性以外の素性を加えても精度がほとんど向上しなかった問題について，新たな素性を加えることで素性がスパースになってしまったことがあげられる．今後は BOW を地名など特に場所特定性が高いものに絞ることや係り受け動詞を同様の移動を表すような動詞でまとめるなどして対処することを検討したい．

文　献

[1] Kinsella, S., Murdock, V. and O'Hare, N.: "I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets, *Proceedings of SMUC*, pp. 61–68 (2011).

[2] 伊川洋平，榎　美紀，立堀道昭：マイクロブログのメッセージを用いた発信場所推定，*Proceedings of DEIM* (2012).

[3] Wing, B. P. and Baldridge, J.: Simple Supervised Document Geolocation with Geodesic Grids, *Proceedings of ACL*, pp. 955–964 (2011).

[4] Roller, S., Speriosu, M., Rallapalli, S., Wing, B. and Baldridge, J.: Supervised Text-based Geolocation Using Language Models on an Adaptive Grid, *Proceedings of ACL*, pp. 1500–1510 (2012).

[5] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J. and Tsioutsiouliklis, K.: Discovering Geographical Topics in the Twitter Stream, *Proceedings of WWW*, pp. 769–778 (2012).

[6] Matsushima, S., Shimizu, N., Yoshida, K., Ninomiya, T.

and Nakagawa, H.: Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class., *Proceedings of SIAM*, pp. 303–314 (2010).

[7] Han, B., Cook, P. and Baldwin, T.: Text-Based Twitter User Geolocation Prediction, *Journal of Artificial Intelligence Research*, Vol. 49, pp. 451–500 (2014).