文脈情報を考慮した旅行中ツイートの判別

狩野 竜示　谷口 元樹　根本 啓一　大西 健司　大熊 智子
富士ゼロックス株式会社
{kano.ryuji, motoki.taniguchi, keiichi.nemoto, takeshi.onishi, ohkuma.tomoko}@fujixerox.co.jp

1 はじめに

1.1 背景と課題

近年，地方都市の過疎化や外国人観光客の増加に伴い，国が交付する地方活性化のための予算は年々上昇し，観光誘致を市政として掲げる自治体も上昇傾向にある．2015 年から日本遺産の認定が始まり，広域観光周遊ルートの形成や，地方空港への LCC の新規就航が活発化している [1]．しかしながら，多くの地方自治体は，どのような方法で観光誘致を行うべきかの知見に欠けている．そうした知見の把握には，どの観光地に需要があるかを分析する技術が必要である．

ソーシャルメディアサービス (SNS) の発展に伴い，製品・サービス・組織等に対する評価情報を大量に発信するようになった．このため，多くの企業が SNS の情報を元に自社製品やサービスの評判分析を行っている．なかでも，Twitter [1] [2]では多くのユーザーが現在の行動や状況を記した文章（ツイート）を投稿しており， その中には，旅行中である事を示すツイートも多く含まれている．観光地の需要の分析には，こうした旅行に関連した（旅行中）ツイートの抽出が必要である．

先行研究 [2, 3, 4, 5] では，緯度経度情報（ジオタグ）や，旅行中であることを表す定形表現をもとに，旅行中ツイートを抽出している．しかし，こうした手法では手がかりの有無に依存するため，断片的な旅行中ツイートしか抽出できない．本来，旅行中のユーザーは継続的に旅行中ツイートを投稿しており，このツイートの中には単体では旅行中とは判断がつかないが，前後の文脈から旅行中ツイートであると判断できるものも少なくない．こうした一連の旅行中ツイートをなるべく多く抽出することで，旅行者の需要をより詳細に分析できると考えられる．

[1]Twitter，http://twitter.com/
[2]Twitter，及び，ツイートはトゥイッター　インコーポレイテッドの登録商標です．

1.2 目的

本研究の目的は，旅行中ツイートコーパスの作成および，旅行中ツイートの抽出モデルの構築・評価である．

旅行中ツイートコーパスの作成においては、収集したツイートに対して単一ラベルと一連ラベルの二種類のラベルを付与する．単一ラベルはツイート単体で旅行中と判断されるもの，一連ラベルは前後のツイートも考慮して旅行中と判断されるものである．作成したコーパスを分析することにより，本研究では既存研究よりも多くの旅行中ツイートを抽出対象としているすることを示す．

作成したコーパスを用いて，単一ラベルと一連ラベルを推定する機械学習モデルの構築と評価を行う．その結果，単一ラベルよりも一連ラベルの判別精度が高い事を示す．また，一連ラベルの推定では前後の文脈情報が精度向上に寄与することを示す．

2 関連研究

Twitter ユーザーの旅行や観光の行動を分析する研究は多く存在する [2, 3, 4, 5]． ユーザーのツイートには旅行以外の内容に関連するものもあるため，旅行に関連した（旅行中）ツイートかどうかを分類する必要がある． 旅行中ツイートを抽出する方法は位置情報を手がかりに用いる手法 [2, 3] と，本文情報を用いる手法 [4, 5] の 2 種類に分けられる．

2.1 位置情報を手がかりとする手法

Twitter において，ユーザーは緯度経度情報（ジオタグ）をつけてツイートを投稿することができる．[2] ではジオタグから人の移動モデルを構築し，またテキストの情報を用いることで移動目的が観光であるかど

うかをクラスタリングしている．[3] では観光地周辺のジオタグが付与されたツイートのうち，本文に「なう」を含むものを旅行中ツイートとして収集している．

ジオタグ付きツイートはツイート全体の約 1 %であると言われており [6]，分析対象となるツイートが非常に少なくなってしまうことが課題である．

2.2 本文情報を手がかりとして用いる手法

[4, 5] では手がかりとなる表現を本文に含むツイートを収集することで，位置情報を手がかりとする手法の課題である分析対象ツイートの網羅性を解決しようとしている．[4] では，観光スポット名を本文に含むツイートのうち，instagram [3]や swarm [4]などのサービスの URL が含まれるものを旅行中ツイートとしている．[5] では,「阿波踊り」,「通天閣」などの地域を連想しやすい単語や「なう」,「楽しかった」などの旅行中に用いられやすい表現を含むものを旅行中ツイートとしている．本文情報を使うことで網羅性は高まるものの，旅行に直接関連した特定の表現が出現するツイートだけを対象としている．

本研究ではジオタグ付きツイートを手がかりにしつつも，表現によらず，その前後のツイート全体から抽出を行っている点が異なる．

3 データ

3.1 ツイート収集方法

旅行中ツイート判別用のコーパスは以下のように作成した．表 1 に記した観光地 23 件を中心に，半径 1km 以内で得られたジオタグ付ツイートを収集した．そこで得られた 846 ユーザーのジオタグ付ツイートを基点として，同ユーザーが前後三日間につぶやいたツイートを収集した．これは，観光地においてジオタグ付ツイートを投稿したユーザーは前後に旅行中ツイートをしていると期待しての事である．収集した期間は 2015 年 9 月から 2016 年 9 月の一年間であり，収集したツイートの内，計 30935 ツイートにラベル付けを行った．

表 1: 学習用データ収集に利用した観光地リスト

観光地	緯度	経度
美ら海水族館	26.694	127.877
壱岐島	33.796	129.715
ハウステンボス	33.086	129.788
太宰府天満宮	33.520	130.537
高千穂	32.744	131.319
厳島神社	34.288	132.327
出雲大社	35.400	132.686
姫路城	34.839	134.694
USJ	34.666	135.433
金閣	35.039	135.729
伏見稲荷大社	34.968	135.778
伊勢神宮	34.456	136.726
白川郷	36.257	136.857
富士急ハイランド	35.487	138.780
高尾山	35.626	139.244
鶴岡八幡宮	35.326	139.556
日光東照宮	36.757	139.600
皇居	35.686	139.753
ディズニーランド	35.630	139.882
牛久大仏	35.983	140.220
函館山	41.759	140.704
松島	38.378	141.073
中尊寺	39.002	141.102

3.2 旅行中ラベル

前述した 30935 件のツイートに二種類のラベル付けを行った．単一ラベルと一連ラベルである．単一ラベルは，ツイート単体から旅行中と判定できるツイートに付けられたラベルであり，一連ラベルは前後の文脈を含めて旅行中と判定できるツイートに付けられたラベルである．表 2 にあるように，単一ラベルが付けられたツイート，付けられなかったツイートはそれぞれ 3329 件，27606 件であった．一連ラベルの場合はそれぞれ 10405 件，20530 件であった．表 3 に二種類のラベルの例を載せる．ここで挙げられている例,「ご飯おいしい」,「雨降ってきた」等のツイート単体からは旅行中であるかは判断できないが，前後の文脈からは判断が可能である．このため，この例には一連ラベルは付与されるが，単一ラベルは付与されない．一連ラベ

表 2: ラベルデータ数

ラベル名＼データ数	ラベルあり	ラベルなし
単一	3329	27606
一連	10405	20530

[3]https://www.instagram.com
[4]https://www.swarmapp.com

表 3: ツイートとラベル例

投稿時刻	ツイート	単一	一連
17/08/07 08:00	今から京都行く	1	1
17/08/07 12:00	ご飯おいしい	0	1
17/08/07 14:00	雨降ってきた	0	1
17/08/08 12:00	お寺キレイ	1	1
17/08/08 16:00	新幹線で帰る	1	1
17/08/09 08:00	今日から仕事	0	0
17/08/09 12:00	昼休み	0	0

ルの定義は，旅行中で現地で起こった事，行った事に関係したツイートとする．そのため，旅行中に旅行と無関係なニュース等について言及しているツイートには，ラベルを付与しなかった．単一ラベルは一連ラベルである事の必要条件であるため，単一ラベルは，一連ラベルに内包される．

3.3 データの統計情報

表 2 にあるように，一連ラベルが付けられたツイートは，単一ラベルよりも 3 倍多い．この事から，従来のように単一ツイートから旅行中ツイートか否かを判断する手法では，旅行情報に多くの見落としが存在していた事がわかる．

ジオタグ付きツイートをした 846 ユーザー中，前後三日間で常時ジオタグ付ツイートをしているユーザーは 8.5 %であり，また，30935 ツイート中ジオタグが付けられたものは 18.0 %であった．観光地にジオタグを付けてツイートするユーザーは前後三日間のツイートにおいて，通常より多くジオタグを付けてツイートする事が判明した．ただし，全体の割合から見ると，ツイート全体を十分に網羅しているとは言えないため，ジオタグが付与されていないツイートにも目を向ける必要がある．

ジオタグ付ツイートをランダムで 1000 件サンプリングして調査を行ったところ，ジオタグ付ツイートは「I'm at 成田国際空港 in 成田市, JP https://t.co/u0sI3H3aA7aL」のように，swarmapp 等のアプリから出力される定型文のみで占められたものが全体の 6 割以上 (626/1000) を占める事が判明した．なお，このような定型文のみのツイートは，予め除去してあるため，上記 30935 ツイートには含まれていない．これらの点から，単にジオタグ付ツイートを旅行中ツイートとしてモデルを学習させる事は難しく，また，同ツイートに旅行中の情報が多く含まれているとは言いがたい．

4 手法

本研究では，3 つのモデルを評価した．単一ツイートの素性を用いて単一ラベルを予測するモデル，単一ツイートの素性を用いて一連ラベルを予測するモデル，複数ツイートの素性を用いて一連ラベルを予測するモデルである．単一ラベルは元々人手でラベル付けされた際，前後の文脈が考慮されていないため，複数ツイートの素性を用いて単一ラベルを予測するモデルの構築は行わなかった．

全てのモデルにおいて，形態素解析には MeCab[5] を使用し，学習モデルには SVM を用いた，ツイートの素性には unigram の BoW(Bag of words) を用いた．複数ツイートを素性として用いるモデルにおいては，前後のツイートの BoW を当該ツイートの BoW の前後に結合したものを素性とした．このため，素性の次元は他のモデルと比べて 3 倍となった．また，前後のツイートが存在しない場合は，該当する位置の BoW をゼロベクトルとした．BoW の次元はそれぞれ 50000 とし，高頻出語上位 50000 語のみを集計した．カーネルは線形カーネルを利用し，実装は scikit-learn[6] で行った．

コーパス中の 846 ユーザーを 9:1 に分割し，前者のツイートを旅行中ツイート判別のための学習用データ，後者のツイートを精度検証用のテストデータとした．ただし，表 2 から分かるように，単一ラベルと一連ラベルではラベルありデータとラベルなしデータの比率が異なる．単一ラベルが付けられたツイートは 1 割強であるが，一連ラベルは約 3 割である．この事が学習精度に影響を与えないようにするため，学習データの正例と負例を同数とした．すなわち，単一ラベルの学習においては，正例負例ともに，3329*0.9=2997 件，一連ラベルの学習においては，10405*0.9=9365 件とした．

5 結果

5.1 ラベルの違いによる判別精度比較

表 4 に，単一ラベルを判別したモデルと一連ラベルを判別したモデルの精度を載せる．一連ラベルの判

[5] http://taku910.github.io/mecab/
[6] http://scikit-learn.org/

別精度はより高い結果を示している．これは，機械学習の判別において，文脈情報を考慮したラベル付けを行った方が，より高精度に旅行中ツイートを取得できる事を意味している．

5.2 文脈情報の有無による判別精度比較

表 4 下段より，単一ツイートのみを素性とした文脈なしモデルと，前後のツイートを素性とした文脈ありモデルでは，後者がより高精度となった．これはツイートを予測する際，前後のツイート情報を参照する事でより高精度に旅行者ツイートの判別ができる事を示している．

表 4: 各手法における精度（F 値）

ラベル＼モデル	文脈なし	文脈あり
単一ラベル	0.396	-
一連ラベル	0.517	0.546

6 おわりに

本研究には，2 つの目的があった．1 つは時系列に沿った一連のツイートデータに旅行中ツイートのラベル付けを行い，旅行中ツイートコーパスを作成する事である．このラベル付け方法では，単一ツイート毎に旅行中ツイートを判別する既存手法より，3 倍ほど多くの旅行中ツイートが得られる事が判明した．

もう 1 つの目的は，旅行中ツイートを判別するモデルの構築とその評価である．まず我々は一連ラベルの判別と，単一ラベルの判別精度の比較を行った．結果，一連ラベルを判別ラベルとして用いる方が，より高い精度が得られる事が判明した．また一連ラベルの予測において，単一ツイートを素性として用いるモデルと，前後のツイートをも素性として用いるモデルを比較した．結果，後者の方が 3 %ほど高精度となり，文脈情報による予測精度の向上が確認された．

本研究で提示したモデルの応用例は以下の二種類を考えている．1 つは，本モデルを利用して判別された旅行中ツイートに出現する単語を分析する事である．これによって，旅行者が具体的に言及している事柄の特定が可能となる．もう 1 つは，性別，年代，趣味判定等の，プロファイル判定技術と組み合わせる事である．これによって，旅行者のプロフィール分布の特定および，特定の観光地を好むユーザー層の特定が可能になる．このように，ユーザと観光資源，評価などの関係性をモデル化することで，観光需要の把握や，観光誘致に向けた施策立案などへの活用を目指していく．

参考文献

[1] 国土交通省, 観光白書平成 28 年度版, http://www.mlit.go.jp/common/001149499.pdf

[2] 前田高志ニコラス, 吉田光男, 鳥海不二夫, 大橋弘忠. 2015. Twitter 位置情報・テキスト情報を用いた人の移動モデル構築と観光地推薦手法の提案. 人工知能学会 第 9 回データ指向構成マイニングとシミュレーション研究会 (SIG-DOCMAS)

[3] 中嶋勇人，新妻弘崇，太田学. 2013. 位置情報付きツイートを利用した観光ルート推薦. 情報処理学会研究報告，pp.1-6，2013.

[4] 新井 晃平, 新妻 弘崇, 太田 学. 2015. Twitter を利用した観光ルート推薦の一手法. 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM2015)

[5] 小原 基季, 森田 和宏, 泓田 正雄, 青江 順一. 2015. Twitter 本文を用いた観光情報抽出及び分析システムの構築. 人工知能学会全国大会. 4M1-4

[6] Umashanthi Pavalanathan and Jacob Eisenstein. Confounds and Consequences in Geotagged Twitter Data. 2015. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 21382148.