CRF を用いた Twitter からのコロナ後の旅行意向の抽出

Extracting Post-Corona Travel Intentions from Twitter Using CRF

丸　照正
Terumasa Maru
広島市立大学　言語音声メディア工学研究室
Language and Speech Media Engineering Laboratory, Hiroshima-City University

概要 新型コロナウイルスの影響によって，訪日外国人の数は大きく落ち込み，日本人の海外旅行 も困難な状況にある．日本政府観光局（JNTO）の統計によると，2021 年 4 月の訪日外国人数(推定)は 10,900 人であった．2020 年 4 月の 2,917 人と比較すると 273.7%の増加であるが，2019 年 4 月の 293 万人と比較すると-99.6%の減少である．しかし，コロナが落ち着けば旅行したいという意欲は存在した．そこで，本研究では Twitter を活用し，どのような旅行を，誰と，どこで，何を体験（経験）したいのかを抽出する手法を提案する．

1　はじめに

新型コロナウイルスの影響によって，2020 東京オリンピックは 1 年延期となり，緊急事態宣言下で無観客開催となった．オリンピック開催を受けて，多くの訪日外国人や日本人の観戦旅行などが期待できたが，無観客開催となったため観光特需は起こらなかった.2020 年夏には GOTO トラベルキャンペーンを政府が打ち出し，観光業界の経済落ち込みを回復させようとしたが，コロナがまた蔓延したため数か月で中止となった．経済と観光業の回復をしよう 7 月から全国の旅行キャンペーンを政府は打ち出す予定だったがコロナの波が再び訪れ延期となった.現在は,県民限定の旅行施策をしているが第 7 派が訪れ自由に旅行が出来にくい状況が続いている．

日本政府観光局（JNTO）の統計によると，2021 年 4 月の訪日外国人数(推定)は 10,900 人であった．2020 年 4 月の 2,917 人と比較すると 273.7%の増加であるが，2019 年 4 月の 293 万人と比較すると-99.6%の減少である．このように，新型コロナウイルスの影響によって，訪日外国人数は大きく落ち込み，日本人の海外旅行も困難な状況にある．

しかし，新型コロナウイルスのまん延が落ち着いたら，旅行したいという意欲は存在する．そこで，本研究では，Twitter を活用し，どのような旅行を，誰と，どこで，何を体験（経験）したいのかを抽出する手法を提案する．

2　関連研究

Twitter を用いた研究として以下のようなものが挙げられる．

伊神ら[1]は，効果的な観光マーケティングのための SNS 分析に着目し，地域観光の SNS データ利活用推進のための取り組みの一環として，Twitter のテキスト分析を行っている．キーワード「飛騨」を含むツイートを収集し，投稿内容に対してテキスト分析を実施している．

新井ら[2]は，観光ルート推薦行うために,Twitter から実際の観光体験を収集し,観光を「食事」，「景観」，「行動」，「土産」に分類し，それらの結果を用いて観光ルートを推薦する手法を提案している．

鈴木ら[3]は，各地域の観光協会の Twitter アカウントを対象に，フォロワのユーザプロフィールを分析し，フォロワの関心の対象から各地域の特徴付けを行う研究をしている．

石野ら[4]は，旅行前の旅行計画者の行動をモデル化することを目的とし，Twitter に投稿されたツイートから，旅行を計画中のツイートと，旅行中のツイートを自動で判定する手法を提案している．

李ら[5]は，Twitter データに基づいた Over Tourism に対する要因の分析と解決方法の考察をしている．

また,CRF を用いた研究として,石野ら[6]は，東日本大震災のような大災害が起きた時被災者の避難経路や救援物資の配送に利用可能な経路の情報は非常に重要な情報であると考え，東日本大震災情報に関連する Twitter のデータから，機械学習を使用して，ユーザの行動経路を抽出する手法を提案している．

3　分析方法

3.1　実験データ

本研究では，Twitter を用いて行う．そのツイートデータは，2021 年 8 月 24 日から 2021 年 12 月 19 日に収集したツイートで，リツイートを省いた 4674 ツイートである．

3.2　Conditional random field について

本研究は機械学習の手法として Conditional random field/条件付き確率場（以下，CRF）を用いて解析していく．CRF とは，無向グラフにより表現される確率的グラフィカルモデルの一つであり，識別モデルである．形態素解析ライブラリの MeCab にも使われている

3.3　正解データの生成

タグの付与をする必要があるため，表 1 の要領で，「旅行形態」，「行き先」，「同伴者」，「経験」にいついてタグを付与する．タグは判別がつくように〈〉で囲い，タグの最後を表すため〈〉の中に「/」を入れ，タグの間の文字を取得できるように処理する．表 1 にタグの種類と説明を示す．

表 1：タグの種類と説明

種類	タグ	説明
形態	〈form〉	どのような旅行をしたいか，○○旅行につける
		例）新婚旅行，卒業旅行，海外旅行など
		〈form〉新婚旅行〈/form〉，〈form〉卒業旅行〈/form〉
		○○が地名だった場合は行き先タグ〈destination〉をつける
		（そのほかのタグに該当する場合も同様）
		例）沖縄旅行：〈destination〉沖縄〈/destination〉旅行
		例）温泉旅行：＜experience＞温泉＜/experience＞旅行
行き先	〈destination〉	どこへ旅行したいか
		例）ハワイ，広島，TDL（ディズニーランド），ヨーロッパ など
同行者	〈toge〉	誰と旅行したいか
		例）家族，友人，親，子どもなど
体験	〈experience〉	何を体験（経験）したいか
		例）ダイビング，登山，温泉，ラーメンなど

3.4 分析手順

① Twitter API を用いて「コロナ後　旅行」の検索ワードでリツイートを省いて，ツイート収集
② 正解データに該当するツイートを手動で抽出し，正解データの作成
③ 正解データと処理していないデータをランダムに混ぜる
④ 混ぜたツイート内にある空白の処理を行う
⑤ ④のツイート内のタグ情報を読み取り，形態素解析を行う
⑥ ⑤の形態素解析を行ったものにタグの有無とはじまり，途中を示す情報である B-I-O を付与する
⑦ ⑥のツイートについて CRF モデルを作成し学習を行う

学習について 8 割を学習データ，2 割をテストデータとする．CRF モデルの作成に使用する情報として，「単語」，「タイプ」，「品詞」，「品詞細分類」が挙げられる．ここでの「タイプ」とは形態素解析された単語のタイプで，ひらがな，漢字，カタカナ，記号を判別するものである．「品詞細分類」は，品詞をさらに細かく分類したもので「広島」であれば「固有名詞,地域,一般」の情報が含まれる．

⑥の B-I-O について，固有表現抽出として，本研究では B-I-O 方式を採用する．タグのはじまりに B（Begin），形態素解析で同じタグであるのに分かれてしまったものに I（Inside），タグが付いていないものには O（Outside）がつけられ，タグの有無と，種類がわかるようになる．その B-I-O のタグについての説明を図 1 に示す．

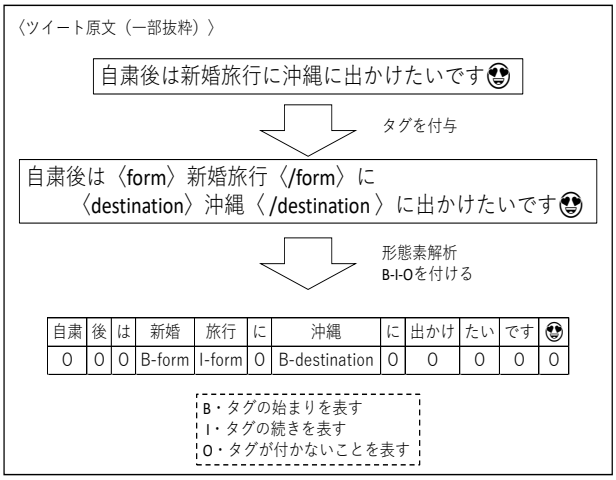

図 1：B-I-O の付け方説明

タグの付与は人手で行い，その後の形態素解析と B-I-O の付与は自動的に行っている．

本研究では，学習に細分類情報を何番目まで使用するかと，参照前後単語数に着目する.その考え方を図 2 に示す.

参照前後単語数	4	3	2	1	0	1	2	3	4
原文	**新婚**	**旅行**	**に**	**沖縄**	**に**	**出かけ**	**たい**	**です**	**。**
単語	**新婚**	**旅行**	**に**	**沖縄**	**に**	**出かけ**	**たい**	**です**	**。**
タイプ	OTHER	OTHER	HIRAG	OTHER	HIRAG	OTHER-HIRAG	HIRAG	HIRAG	OTHER
品詞	名詞	名詞	助詞	名詞	助詞	動詞	助動詞	助動詞	記号
細分類1	一般	サ変接続	格助詞	固有名詞	格助詞	自立	*	*	句点
細分類2	*	*	一般	地域	一般	*	*	*	*
細分類3	*	*	*	一般	*	*	*	*	*

図 2：参照前後単語数と細分類情報数の考え方

4 実験・結果

3.4 の分析手順にそって実験を行った．その実験結果を図 1 に示す.

また，この数値は，それぞれのタグの F 値を平均した micro 平均を表している．今回一番値が良かったのは，細分類情報を 3 番目まで付与し，参照前後単語数が 3 の時だった.

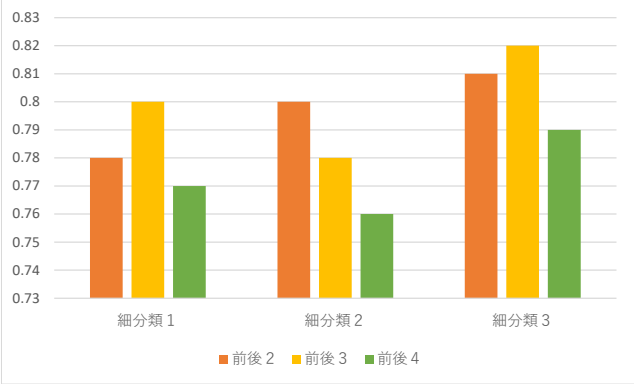

図 3：細分類情報と参照単語数の変

5 LSTM の検討実験

LSTM（Long Short Term Memory）を用いた検討実験を行った.それぞれのタグの分類問題として LSTM を用いた.

CRF と同様に B-I-O の分類した結果を表 2 に示す.

表 2：B-I-O の LSTM 分類結果

	precision	recall	f1-score
B-destination	0.04	0.20	0.07
I-destination	0.25	0.13	0.17
B-experience	0.12	0.30	0.17
I-experience	0.12	0.20	0.15
B-form	0.07	0.40	0.12
I-form	0.08	0.28	0.13

次に,B-I の区別をなくした結果を表 3 に示す.

表 3：B-I の区別をなくした実験結果

	precision	recall	f1-score
destination	0.93	0.82	0.87
experience	0.90	0.95	0.93
form	0.87	0.92	0.89

今回の LSTM の実験では,予備実験であまり数値がよくなかったため,train データにタグのついているツイートのみを利用している.train 内にあるタグのついているツイートは CRF の train 内についているツイートと同じである.

6　考察

それぞれのタグにより数値に偏りがみられた．細分類情報３つ，参照前後単語数３の時の結果を表 4 に示す.

表 4：参照前後単語数３と細分類情報数３

	precision	recall	f1-score
B-destination	1.00	0.28	0.44
I-destination	0.75	0.75	0.75
B-experience	1.00	0.50	0.67
I-experience	0.78	0.70	0.74
B-form	1.00	0.96	0.98
I-form	1.00	1.00	1.00

「form」が一番高くなっており，当てやすいことがわかる．今回，「〇〇旅行」にタグを付与したため，予測することが容易であったと思われる．「experience」は，経験することにつけられており，「〇〇巡り」は「form」の時と同様に予測しやすいが，そのほかのものに関しては経験することは多岐にわたるため当てにくい予測になっていると考える．「destination」は地名や国名などの行先に付与している．このタグについても予測するものがより多岐にわたるため世当てにくいものとなっていると考える．また，同伴者タグは固有名詞なども入ってくるため抽出することができなかった.

LSTM の実験では,CRF 同様に B-I-O の区別で実験したものはあまりいい結果が得られなかった.これは分類が多くなると,その一つ一つの正解データが少なくなるため当てるのが難しくなったと考えられる.そのため,B-I の区別をなくした実験では値はよくなった.タグごとの正解データが多くなったため当てやすくなったと考える.しかし,B-I の区別をなくしているため,形態素解析を行った際に 2 つに分かれてしまうもの（I のつくもの）だけでは,何なのか判別がつかないため抽出はできていないのと同じだと考える.

7　まとめ

今回，「コロナ後　旅行」の検索ワードでツイートを取集したため，多くのツイートが集まらなかった．また，タグを付与することのできるツイートが全体の約７％だったため，学習が十分に足りなかったと予測される．ツイートを多く収集するための検索ワードの検討や，タグの種類の検討などが必要だと考える．コロナへの対応は日々変わっており，「with コロナ」の生活様式になっている．その対応や生活様式に合わせた研究や解析をしていくべきだと思う.

8　参考文献

[1] 伊神 花織，浦田 真由，遠藤 守，安田 孝美”岐阜県飛騨市の観光推進に向けた Twitter のテキスト分析”，観光情報学会第 22 回研究発表会講演論文集，pp. 5-8（2021）

[2] 新井 晃平, 新妻 弘崇,太田 学, “Twitter を利用した観光ルート推薦の一手法”, DEIM Forum（2015）

[3] 鈴木祥平，倉田陽平，Twitter のユーザプロフィールを用いた観光地の特徴分析，観光情報学会誌「観光と情報」，第 13 巻,第 1 号，pp. 39−52, 2017.

[4] 石野亜耶，難波英嗣，竹澤寿幸，Twitter を利用した旅行計画者の行動分析，観光情報学会 第 16 回研究発表会講演論文集，2017.

[5] 李昭知，中嶋卓雄，Twitter データに基づいた Over Tourism に対する要因の分析と解決方法の考察，観光情報学会 第 20 回研究発表会 講演論文集，pp. 41−43, 2019.

[6] 石野亜耶，小田原周平，難波英嗣，竹澤寿幸：”Twitter からの被災時の行動経路の自動抽出および可視化”，言語処理学会第 18 回年次大会発表論文集，pp. 907-910（2012）