# CRF を用いた Twitter からのコロナ後の旅行意向の抽出

丸照正†　石野亜耶‡　目良和也†　竹澤寿幸†
†広島市立大学　‡広島経済大学

キーワード：Twitter，旅行意向，コロナ

【目的】

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．

【分析方法】

＜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．

＜Conditional random field（条件付き確率場）について＞

本研究は機械学習の手法として，Conditional random field（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．

＜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 |

＜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．

【実験，結果】

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し，上記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．

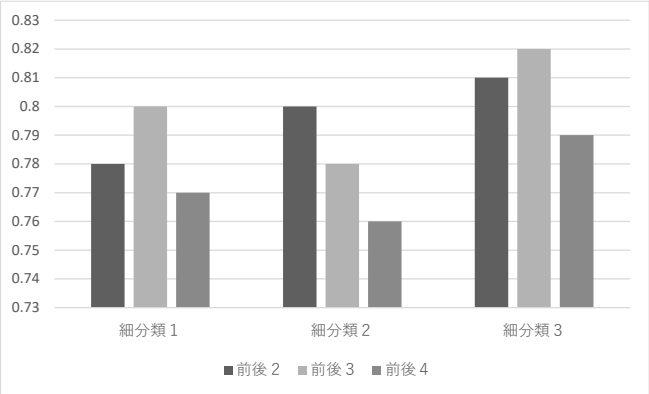


**図 1：細分類情報と参照単語数の変**

【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 2 に示す．「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，同伴者タグは抽出することができなかった．

**表 2：細分類情報 3，参照前後単語数 3 の結果**

| | 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 |

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．