Twitter を対象とした観光情報の現地性判断

嶋田 和孝　九州工業大学 大学院情報工学研究院 知能情報工学研究系
井上 俊右　九州工業大学 情報工学部 知能情報工学科
遠藤 勉　九州工業大学 大学院情報工学研究院 知能情報工学研究系

キーワード：観光情報抽出，Twitter，現地性判断

１．はじめに

現在，さまざまな地域の基幹産業の一つとして観光業が注目されている．その環境下で，Web に存在する観光情報は重要な役割を持っている[1]．我々は，Twitter を対象とした観光情報の分析システムの構築を進めている[2]．提案システムでは，市などが整備した観光情報のポータルサイトから抽出した観光名所やイベントに関連する語を利用して，Twitter から観光情報を抽出する．しかしながら，関連語が含まれるからといって，必ずしも抽出された情報が分析に必要なものとは限らない．例えば，「来週<<観光地名>>に行くつもり．」という Tweet（Twitter における投稿文）は，観光の評判などを分析システムにとっては必ずしも適切な入力ではない．

本研究では，関連語によって抽出された Tweet が，観光情報分析システムの入力として相応しいかどうかを判別することをタスクとする．具体的には，その Tweet が観光地でつぶやかれているか，という「現地性判断」を行う．現地性を持った Tweet のみを獲得できれば，ユーザの実際の経験に基づく情報が獲得され，観光情報としての信頼性が向上する．また，ユーザの現在地を特定できれば，観光地での行動分析などにも有用である．提案手法では，ルールと機械学習を組み合わせ，精度の高い現地性判断を実現させる．

２．提案手法

提案手法では，まずルールベースのフィルタリングによって明らかなノイズを除去し，その後，機械学習によって現地性を判別する．

２．１　ルールに基づくフィルタリング

Twitter 全体の Tweet を考えると，その多くは現地性がない Tweet である．そのため，事前にルールに基づいて，明らかに現地性のない Tweet を取り除く必要がある．このフィルタリングには以下のようなルールを用いる．

- 言語情報：予定を表す「明日」や「いつか」のような語，推定を表す「だろう」や「みたい」などの語，伝聞を表す「らしい」のような語，Twitter 独自の対話構造を表す「@」や「RT」などを含む Tweet は実体験でない可能性が高いので取り除く．
- 時間帯：深夜帯の Tweet は雑談が多数を占める．また，深夜に観光をしている可能性は極めて低いので，23 時～午前 3 時までの Tweet を取り除く．
- Tweet 長：事前実験により，100 文字以上の Tweet は現地性のないもの（宣伝など）が多い．そのため，100 文字以上の Tweet を除外する．
- 名詞数：名詞が多く含まれる Tweet は企業の宣伝などであることが多いため取り除く．

一方で，単純にルールを適用すると現地性のある Tweet も削除される可能性がある．そこで，現

地性の高い単語（「～に到着」や「～を見ている」）などの語が存在する場合は，その Tweet を除外しないものとする．

２．２　機械学習による現地性判断

フィルタリング処理を通過した Tweet に対して，機械学習を用いてその現地性を判断する．機械学習には，Support Vector Machines (SVM) を用いる．SVM の素性としては，以下を用いる．

- Bag of words (BOW)：Tweet 中に出現する単語そのもの
- 言語的特徴：形容詞の時制，動詞の数，名詞の数，地名の有無，「到着」や「～なう」などの特定の単語の有無など．
- 時間情報と時間帯：Tweet された時間や「朝」「昼」「夜」などの時間帯
- Twitter 特有の特徴：各 Tweet の文の長さや RT の有無など．

３．実験

Twitter からランダムに抽出した 2886 tweets に対して，人手で現地性の有無をタグ付けした．その結果，509 tweets が現地性有り，2377 tweets が現地性無しと判断された．このデータを用いて実験を行う．

まず，フィルタリングの精度について述べる．実験データに提案手法を適用したところ，2886 tweets から 1148 tweets まで削減された．このうち，現地性有りが 486 tweets であり，現地性無しが 662 tweets であった．このフィルタリング処理によって 23 tweets が誤って削除されたことになり，精度としては 95.5% (486/509)である．このフィルタリングの精度は現地性判断の精度に大きく影響するため，より精度の高い手法へと今後改良が必要である．一方で，現地性無しの Tweet を 30%程度(662/2377)まで圧縮できており，フィルタリングとして十分機能している．

フィルタリングされたものについて，SVM を適用し，10 分割交差検定で評価した．その結果を表 1 に示す．実験結果より，すべての素性を組み合わせた場合がもっとも精度が高く，提案手法の有効性が確認された．また，フィルタリングを行わなかった場合(機械学習の素性はすべて利用)の精度は 75.0%，再現率は 58.2%であった．表 1 のフィルタリング後にすべての素性を利用した場合と比較すると精度で 5.5%，再現率では 6.8%の差があることになる．この結果から，フィルタリングを適用することの有効性も確認された．

表 1. 実験結果

素性	精度	再現率
BOW	80.2	62.3
BOW+形容詞時制	80.3	62.3
BOW+動詞数	80.3	62.3
BOW+名詞数	80.2	62.1
BOW+地名	79.2	62.9
BOW+特定の単語	79.4	64.6
BOW+時間情報	80.4	63.1
BOW+時間帯	79.9	62.7
BOW+文長	***80.5***	62.5
BOW+RT	80.3	62.3
すべて	***80.5***	***65.0***

４．おわりに

本研究では，観光情報分析システムのための現地性判断についての手法を提案し，評価した．精度は 80%を超えたが，再現率は 65%であり，今後，手法の改善が必要である．

【参考文献】

[1] 斉藤. Web における観光情報の提供と分析. 人工知能学会誌, 26(3):234-239, 2011.

[2] 嶋田, 前田, 井上, 遠藤. 地方都市を対象とした観光情報の分析に向けて. 観光情報学会 第 3 回研究発表会, pp. 8-13, 2011.