

Twitterにおける検索語と関連の高い地名の抽出手法の研究

原 克彬[†] 中山 泰一[†][†]電気通信大学情報・通信工学専攻

1 はじめに

マイクロブログの一つである Twitter の急成長は記憶に新しい。ここ最近では Facebook や LINE などのサービスの台頭もあり、Twitter の勢いはやや落ち着きを見せてきたと言えるだろう。とはいえ、そのユーザ数とデータ量は顕在であり、API も公開されているため頻繁に研究対象として利用されている。

研究で Twitter を利用する場合には、Twitter が保存しているデータを取得し、そのデータのある指標やパターンに基づいて加工してから利用することが多い。Twitter の一投稿（以下ツイート）は、その短い本文量に反して抱えている情報量が非常に多い。それは、設定や投稿の仕方によって多少異なるが、ツイートにはそれぞれ本文の他にツイートした日時やツイートが行われた緯度経度、またツイートが行われた場所の地理情報等、さまざまな付加情報があるからである。このように指標として利用できる情報が多数存在する。その中から、利用者は自身が求めるツイートにあわせて適切な指標を使用する。

指標の使用例として関連の研究を挙げると、Twitter から電車の遅延情報や地域特徴語を抽出するものがわかりやすい。荒井ら [1] の電車の遅延情報を抽出する研究では、本文と日時を指標に、特定の路線に関する最新のツイートを収集し、利用している。伊藤ら [2] の地域特徴語を抽出する研究では、ツイートが行われた緯度経度を利用して特定の地域でツイートされた内容を収集し、利用している。

Twitter における土地関係の指標は主に緯度経度である。しかし、緯度経度情報付きのツイートは数が多くなく、土地に関するツイートのリアルタイムな推移を観測するには不十分だと考えられる。そこで、本研究では、ツイートより地名を抽出したものを新たな指標として提案した。地名抽出の手法を考案、実装し、抽

出された地名とその時の情報を考察した。その結果、地名の検出数が多ければ多いほど、重要度の高い情報と関連がある傾向を観測することができた。また、情報の種類によって各時刻の検出数に差異があることも確認できた。

2 地名の抽出手法

本研究において用いたプログラムは四工程に分かれている。まず、第一工程では特定のキーワードを含む最新の 100 件のツイートを TwitterAPI を用いて取得する。なお、取得数が 100 件なのは一回の取得上限が 100 件となっているからである。第二工程では、取得したツイートに形態素解析処理システム IGO[3] を用いて形態素解析を行う。次の第三工程では、形態素解析から得た単語群を辞書データを参照し、地域として登録されている単語を抽出する。また、一部の地名は一般的な固有名詞として登録されているため、そちらも抽出を行う。今回の実験では辞書データは IPAdic¹ を用いた。この第三工程では同地名の出現回数も集計する。最後の第四工程では、出現回数順に地名をソートして出力する。

3 実験

3.1 実験方法

本実験では特定のキーワードを基に取得したツイートより地名を抽出し、得た地名とその事象の関連がどの程度であるか調べた。また、その情報の推移を観測するため、ある程度時間を空けて、同キーワードで同様にツイートを収集した。関連の判定に関してはその当時のニュースやツイートなどを参考に照合していく。

3.2 結果

2013 年 10 月 16 日のツイートに土砂崩れをキーワードに検索した結果を図 1 と表 1 にまとめた。それによって得られた結果の中で非常に検出数が多かった 2 つの地名を時刻毎のグラフにしたのが図 1、出現回数が多かった 5 つの地名に関する情報をまとめたのが表 1 で

A Study on Extraction Method of Place Names Closely Related to Search Keys in Twitter

Katsuaki Hara[†], Yasuichi Nakayama[†]

[†]Graduate School of Informatics and Engineering, The University of Electro-Communications
182-8585, Chofu, Japan
hara-k@igo.cs.uec.ac.jp

¹IPA コーパスに基づき CRF でパラメータ推定した辞書

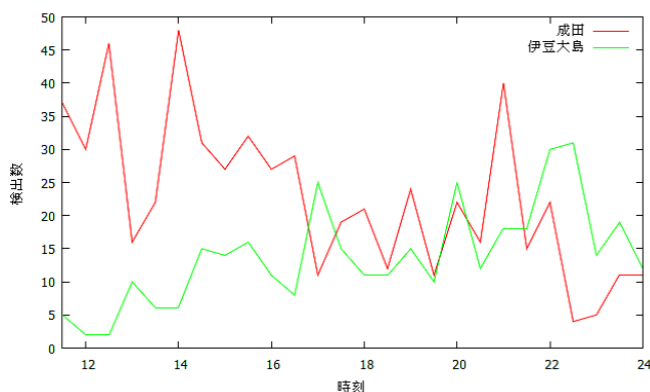


図 1: 成田と伊豆大島に関するツイート数の推移

表 1: 10 月 16 日の土砂崩れに関する一部の情報と関連ツイートを検出した時刻

ニュース	検出時刻
茨城・行方市で土砂崩れ、3 人軽傷	14:00~15:00, 16:30~17:30, 19:00, 21:00~22:30, 24:00
京成成田駅で線路脇の土砂流出 架線切れ運転見合わせ	11:30~24:00
伊豆大島 13 人死亡 50 人連絡取れず、土砂崩れ多発	11:30~24:00
横浜市西区 5 世帯に避難勧告	16:00~16:30, 19:00, 21:00~22:00, 23:00
東京・町田で 1 人死亡 鎌倉で土砂崩れも	13:30~15:30, 21:00

ある。なお、図 1 の伊豆大島の検出数は本システムによって得られた結果ではなく、手動にてまとめたものである。この問題についての詳細は第 5 章にて述べる。

4 考察

図 1 を見ると、成田という地名を含むツイートが非常に多く存在していることがわかるだろう。同日午前 5 時 50 分頃に台風の影響で京成成田駅で土砂崩れにより列車の運行が出来なくなったことが影響していると考えられる。京成線の利用者やニュースや画像を見たユーザがツイートしたり、他のユーザのツイートをリツイートするなど非常に高い注目度を獲得している事がわかる。

グラフを見ると時刻によって検出数にかなり差が生じていることがわかる。成田に関していえば 12:30, 14:00, 21:00 に非常に多く検出されており、反対に 13:00 や 22:30 以降はかなり落ち込んでいる。12:30 頃に多く検出されている理由はおそらく、京成線に関連するツイートが非常に多い中、多くの人がお昼休みに入ったことでリツイート等をするユーザ数が一時的に増加したからと考えられる。また、21:00 頃については、仕事等を終え、帰宅途中に携帯等から情報を得てツイート、ないしは京成線の利用者がツイートした結果、検出数が増加したと考えるのが妥当だろう。ただ、14:00 頃的大幅な増加については考慮したものの、原因を予想することはできなかった。天候の具合を見て、多くの学校や

企業がお昼を過ぎたところで帰宅を促した結果、21:00 頃の現象がこの時間帯にも起きた可能性はある。また、13:00 頃と 22:30 頃に関しては予想が容易で、お昼休み明けであることと、活動しているユーザ数が少ないからだと考えられる。

データの全体を見渡してみると、重要度の高いニュースや情報に関連する地名ほど発生から長い間検出が確認されたり、幾度と検出されたりしている。この傾向は、表 1 からも見取れる。これは、重要度の高い情報ほど、再報道や続報が多くもたらされているため、この結果になったものと考えられる。

5 まとめと今後の展望

本稿では、ツイートより地名を抽出する手法の提案と地名を指標として利用した際の検証を行った。Twitter から抽出した地名を指標とすることで、同じ内容について異なった書き方がしてあるツイートをまとめて、注目度の算出を行うのには有効性を見出せたと考えている。

今回の実験を通して、いくつかの問題が浮き彫りになった。特に大きいのが、伊豆大島のような特定の単語が連続する地名に関するものである。第三章の結果のところでも触れたが、関連するツイートは多数存在していたものの、本手法を用いて地名を抽出した結果の中に伊豆大島の名前は見られなかったのである。これは本システムにて利用している形態素解析器と辞書が原因であることがわかっている。まず、伊豆大島という地名が形態素解析器によって伊豆と大島に分けられてしまい、分割後に辞書を参照すると高確率でどちらも人名と判断されてしまう。また、一部の地名は地名としてではなく一般的な固有名詞として辞書に登録されているため、地名として参照した場合には検出ができない。地名を指標として利用するには正確に抽出を行わなければいけないので、今後の展望としては地名に強い辞書データの作成が挙げられる。

参考文献

- [1] 新井誠也, 平川豊, 大関和夫, “Twitter からの列車遅延情報収集手法の検討”, 情報処理学会研究報告, 情報学基礎研究会報告 Vol. 112, No. 1, pp1-8, 2013 年 9 月.
- [2] 伊藤晶, 荒川豊, 田頭茂明, 福田晃, “Twitter からの地域特徴語の自動抽出に関する一検討”, 第 75 回情報処理学会全国大会 (2013).
- [3] Igo - Java 形態素解析器
<http://igo.sourceforge.jp/>