ビッグデータを用いた災害発生箇所の情報抽出アルゴリズムの開発

愛媛大学　学生会員　堀　太成　　愛媛大学大学院　学生会員　○泉　翔太
東京大学　　正会員　全　邦釘　　愛媛大学大学院　　正会員　　森脇　亮

1.　背景・目的

我が国は世界有数の災害大国といわれており，高い頻度で地震が発生し，2019 年には台風 19 号の上陸により関東・中部地方をはじめとして甚大な被害をもたらした．国としても対策を行っているが，自然災害は人間の予想を上回ることも多く，災害による死亡者数を減少させるに至っていない．ゆえに更なる減災対策が必要不可欠であるといえる．

その対策の 1 つに発災時の初期段階から求められるものとして情報収集が挙げられ，近年ソーシャル・ネットワーキング・サービス(以下 SNS)を活用し役に立った案件が増えている[2)]．特に，SNS の 1 つである Twitter は実際に熊本地震の際，その情報が Twitter へ投稿され，一週間で約 2610 万件となった．この数字は，東日本大震災の時の投稿数の約 23 倍(約 115 万件)に上り，情報伝達および収集の手法の 1 つとなった．

しかし，Twitter 公式アプリケーションを用いて災害のキーワードを入力しツイートを検索すると，防災に有用な情報が含まれるツイートとそうでないツイートが混合して表示されることから，Twitter の公式アプリケーションが災害時の情報取得にあたっては最適ではないといえる．現時点で SNS 上の災害情報を収集・整理し発信するシステムとして D-SUMM (災害状況要約システム[3)]が運用されている．このシステムにより，災害時にも現地に行く前から一定量の情報を取得することができる．しかし，D-SUMM にまとめられた情報の中にも，災害に関係のないツイートが一定数存在するため，このような情報をフィルタリングし除去する必要がある．

そこで本研究では，災害の中でも台風と大雨を対象とし，台風が接近・上陸時と大雨が降った際に，Twitter への投稿を取得し投稿から防災上，有用な情報のみを抽出することを試みた．抽出には文章を精度よく分類できる Deep Learning[4)]をモデルの 1 つである BERT[5)]を用いることで．また抽出した情報をマップ上に表示し，位置情報を同時に見ることができるようにすることを目的とした．

2.　BERT

BERT は自然言語処理を行うための Deep Learning モデルである．BERT は attention モデルを採用したことで知られている Transformer を双方向的に学習させることに適応させたことが特徴に挙げられる．つまり従来の自然言語処理モデルが文字列(文章)を文頭から単方向的に読み取り訓練していたのに対して，BERT は文字列を一度に読み込むことによって，文脈や文脈に応じた単語の意味をより深く読み取ることが出来るようになっている．BERT の学習は，事前学習とファインチューニングの 2 ステップで構成されている．事前学習では用意した大量のテキストデータで学習することによって言語知識を獲得し単語分散表現を計算し，その後分散表現を用いてファインチューニングを行う流れである．本研究では，公開されている事前学習データの京都大学 黒橋・川原・村脇研究室の BERT 日本語 Pretrained モデルを使用した．

3.　研究方法

まずツイートの収集には Twitter API を利用した．Twitter API とはプログラムレベルで Twitter にアクセスし，ツイート情報を収集できるアプリケーションである．収集したツイートは台風 15 号・台風 19 号・台風 21 号が日本に上陸した日からその台風が温帯低気圧に変わるまでに投稿されたものと，2019 年 11 月 22 日，12 月 2 日と 2020 年 1 月 23 日に投稿されたものである．

次に Deep Learning モデルを作成するために，収集したツイートを整形した．Twitter に投稿される文字列には URL やツイート内容に関係のない顔文字等が多く含まれるため精度に期待できない可能性が高いためである．

また，全角数字を半角数字に，大文字のアルファベットを小文字のアルファベットに変換するなどした．整形したツイートを，必要な (台風・大雨に関係のある)ツイートと，不必要なツイートに手作業で分類し，学習データを作成した．その結果，必要なツイートが 1081 ツイート，不必要なツイートが 3262 ツイートとなったが，このデータでそのままモデルを学習させると，ツイートを全て不必要な情報へ分類してしまう恐れがあるため，不必要なツイートの数を必要なツイートの数と同数にするダウンサンプリングを行った．分類する基準として，1 つのツイートの内容のうち，必要な情報が 8 割以上含まれていれば必要なラベルを，2 割以上不必要な情報が含まれていれば不必要なラベルを付与した．この学習データを用いて学習を行い，台風 19 号のツイートを作成したモデルで分類し，必要と分類されたツイートを地図上に表示する．

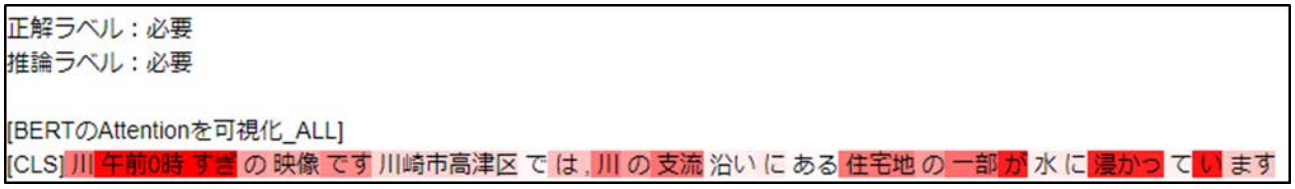

図 1 注目箇所

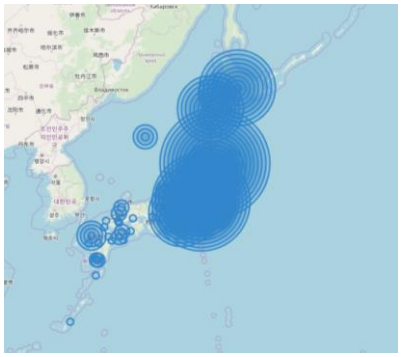

図 2 分類前ツイート

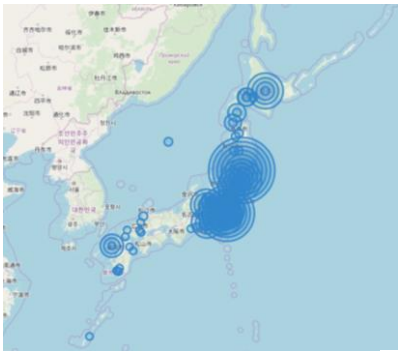

図 3 必要なツイート

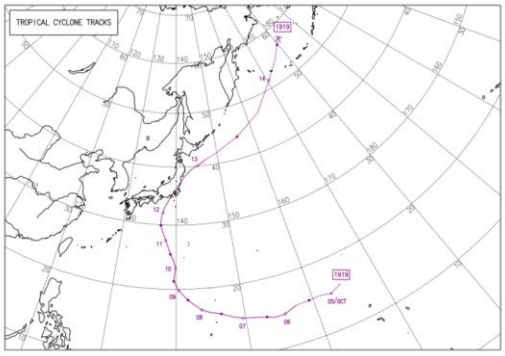

図 4 台風 19 号経路図

[気象庁：https://www.data.jma.go.jp/fcd/yoho/typhoon/route_map/bstv2019.html]

4. 結果・考察

モデルの精度検証は再現率を用いて評価する．作成したモデルの精度検証の結果は真陽性率(再現率)が 0.760，真陰性率が 0.858 であった．真陽性率の結果からは，実際に必要な情報のうちの 76%を実際に分類できているのが分かる．また真陰性率の結果も真陽性率と同様に実際に不必要な情報のうち，85.8%を正確に分類できていることが分かる．この 2 つの値から，必要な情報・不必要な情報のいずれもある程度正確に分けられているといえる．また，図 1 はモデルが注目するほど色が濃くなるように出力し可視化したものであり，伝えたい情報の時間や住宅が水に浸かっている点など，人間が着目する単語に注目していることが分かる．

次に図 2 は 2019 年台風 19 号に関する投稿を作成したモデルで分類を行った結果である．その結果必要な情報と分類されたのが 341 ツイート，不必要な情報と分類されたのが 439 ツイートとなった．そのうちの分類前のツイートと，必要なツイートを地図上に表示したものをそれぞれ図 2，3 で示す．地図上に表示されている円は，円の中心がツイートから抽出した地名の位置であり，1 ツイートにつき 1 つの円をプロットしている．図 2 と 3 を比較すると，図 2 より図 3 の方が明らかにツイート数の減少が見られ，図 4 の台風経路図と比べると，不必要な情報も入っている図 2 の状態では，全国各地で台風 19 号に対して様々な投稿がされているが，必要な情報だけに絞り表示した図 3 では台風 19 号が実際に通った地域で多くツイートされているということが分かる．これにより，作成したモデルが未知のツイート郡に関してもある程度の汎化性能を有することが示された．必要な情報の中でも非常に危険な状況を示すものとそうでないものが含まれているため，今後は今回必要な情報と振り分けたデータの中でもさらに細かく振り分けられるモデルを作成する必要があると考える．