E-015

ニュース取材支援のためのソーシャルメディア分析システム
Social Media Analysis System for News Gathering Support

武井友香† 牧野仁宣† 宮﨑太郎† 住吉英樹† 後藤淳†
Yuka Takei Kiminobu Makino Taro Miyazaki Hideki Sumiyoshi Jun Goto

1. はじめに

Twitter や Facebook などの SNS (Social Networking Service) 上で大量に発信される情報は，ソーシャルビックデータと呼ばれ，放送局にとって有力な情報源の一つである．ソーシャルメディアを常時監視し，事件や事故の現場に居合わせた投稿者から直接情報を得ることで，より迅速に報道することができる[1]．現状では，事件や事故を表現するようなキーワードの組み合わせで Tweet を絞り込み，該当する Tweet を人手によって探している．しかし，大量の Tweet から有益な情報を手動で抽出するには，多大な労力を必要とする．そこで，我々は事件や事故に関する情報を含むニュース性のある Tweet を自動で抽出する手法の研究に取り組んでいる[2]．報道現場において，実際にニュースの取材対象となった Tweet に火事や交通事故などのニュースのカテゴリを表すラベルを付与し，ニューラルネットワークを用いて学習することで，現場で必要とされる情報を自動で抽出，分類することが可能である．しかし，ニュース性ありと抽出した Tweet にも誤りが含まれており，最終的に制作者がチェックする必要がある．

そこで，事件や事故に関する Tweet を制作者に提示しつつ，ニュース取材に役立つと制作者が判断した Tweet と，不要とした Tweet を特定するためのインターフェースを備えた，ソーシャルメディア分析システムを開発した．本システムを使用することで，情報の監視と共に新たな学習データを取得することが期待できる．

現場利用を想定した評価者が約 40 日間にわたりアノテートを実施した結果，75,022 件の学習データを取得できた．これらを追加学習用データとし，新たな学習モデルを作成したところ，システムの性能が向上することを確認した．

2. ニュース性のある Tweet の抽出と提示

2.1 ニューラルネットワークを用いた Tweet 自動抽出

まず，ニュース取材に役立つ Tweet を自動抽出するニューラルネットワークについて述べる．本稿では，宮﨑らの RNN (Recurrent Neural Network)を用いた Tweet 自動抽出手法を用いる[2]． ソーシャルメディアでは，気軽に投稿ができるという特徴から，口語体で書かれることが多く，一般的な形態素解析などの手法を用いるよりも，入力文章を文字単位で扱うことで良い性能を得られるという報告がされている[3]．宮﨑らの手法では，単語分割せずに，文字ごとに順方向と逆方向の 2 つの RNN に入力し，それぞれの RNN の内部状態を結合し，Tweet 全体の意味を表すベクトルを得る．得られたベクトル表現を，2 層の FFNN (Feed-Forward Neural Network)によりニュース取材に役立つかどうか判定する．また，Tweet の入力文の重要部分に重み付けをするアテンションメカニズムと，入力文字列の次の文字を予測するタスクと複数のタスクで学習することで，抽出精度が向上することが確認されている．

本稿では，ニュースに役立つと判定された Tweet に，23 カテゴリに細分化したラベルを付与し学習データとして用いることで，ニュース性のある Tweet の自動抽出と同時にニュースカテゴリも分類する．

2.2 制作者に Tweet を提示するインターフェース

次に，ニュース性のある Tweet を制作者に提示するソーシャルメディア分析システムを図 1 に示す．2.1 節の手法で生成したモデルを用いて，1 日 800 万件程度の Tweet をリアルタイムに解析し，約 0.1%の Tweet を抽出している．

図 1 中の黄色の四角枠内では，1 件の Tweet のテキスト本文・投稿日時・添付画像・GPS 情報（またはシステムが特定した投稿場所）・ニュース性の数値（2.1 節のニューラルネットワークが算出した該当カテゴリの数値）を示している．マルチクラス分類のため，ニュース性のある Tweet は，カテゴリごとに 23 種類のラベルを付与されており，制作者が注目するカテゴリを自由に選択することができる．通常は，それぞれのカテゴリに該当する Tweet 数のバランスを踏まえ，「火事・火災」「列車・交通事故」「気象・災害情報」「その他」の 4 種別に分け，制作者に提示している．本システムでは，ニュースカテゴリ分類とともに，情報通信研究機構が開発した地名データベース[4]を用いて，Tweet 文中の地名表現やランドマーク名から投稿場所を特定している．ニュースカテゴリと都道府県ごとに分類することで，監視対象数を絞ることができ，地域放送局において，重点的に監視することも可能である．具体例を図 2 に示す．

図 1. ソーシャルメディア分析システム

† NHK 放送技術研究所

図 2. ニュースカテゴリ・都道府県に分類提示

図 2 のインターフェースでは，抽出した Tweet をカテゴリ・都道府県別に分類した後，該当する Tweet に含まれる単語の頻度を示すタグクラウド，投稿件数の推移を示す折れ線グラフ，発生した場所を示す地図情報を表示している．

制作者はそれぞれの Tweet がニュース取材に役立つかどうかを判定し，「good」または「bad」ボタンを操作する．このログはシステムに蓄積され，ニュース性判定モデルを更新するための，新たな学習データとして利用できる．

3. 評価実験

今回開発したインターフェースで得られた新たなデータの効果を確認するために，2 つの評価実験を実施した．実験 1 では，報道現場に即した評価データを，実験 2 では，キーワードに関わらず Tweet を解析しているシステムの現状を評価するデータを用意し，モデルの性能を比較する．

3.1 実験条件

初期学習データとして，正例には報道現場で実際にニュース制作に役立つと判定された Tweet 40,170 件，負例にはランダム抽出した Tweet 1,535,702 件を用いた．また，追加学習用のデータとして，開発したインターフェースを用いて 2017 年 6 月，10 月，11 月，12 月中に約 40 日間，1 人の評価者がインターフェースを操作し，75,022 件の追加用の学習データを取得した．正例は 34,605 件，負例は 40,417 件である．なお，本稿で用いるモデルは Tweet のテキストのみからニュース性の有無を判定するため，画像や URL 先の確認が必要な Tweet は判定対象外とした．初期学習データのみを学習したモデルを Baseline とし，新たな学習データを加えて学習したモデルを追加学習モデルとする．

評価データとして，2 種類のデータセットを用意した．1 つ目は，報道現場で使用されている事件や事故に関するキーワード約 210 単語の組み合わせでフィルタした Tweet からランダムに 2,000 件抽出し，1 名の評価者によりそれぞれの Tweet についてニュース制作に役立つかどうかのラベルを付与した（キーワードフィルタあり評価データ）．2 つ目は，2017 年 9 月中の 2 億 4 千万 Tweet の中から，何もフィルタをかけずランダムに 10 万件を選んだ．この中で，各モデルが抽出した Tweet を評価者が正例・負例・不明の 3 値に分類した（キーワードフィルタなし評価データ）．

ニューラルネットワーク実装には Chainer [5]， RNN の実装は LSTM を利用した．中間層のノード数は双方向 RNN が 200，FFNN の層は入力層に近い方から順に，200，100 とした．Baseline，追加学習モデルともに，同一のパラメータを用いている．

3.2 実験結果

表 1 に報道現場に即したキーワードフィルタあり評価データを用いた実験結果を，表 2 にキーワードフィルタなし評価データを用いた実験結果をそれぞれ示す．

表 1　実験結果 1（キーワードフィルタあり）

手法	Precision	Recall	F 値
Baseline	64.09	85.91	73.42
追加学習モデル	78.38	84.27	81.22

表 2　実験結果 2（キーワードフィルタなし）

手法	抽出総数	正例	負例	不明
Baseline	191	60	102	29
追加学習モデル	136	83	35	18

実験結果 1 では，Baseline に比べ，追加学習モデルの方が，F 値で 7.8 ポイント性能が向上しており，本システムを用いて収集した新たな学習データの有効性が確認できた．

実験結果 2 では，Baseline の適合率が 31%（60/191）であった．これは，現場のキーワードフィルタを経て取得した正例を学習している Baseline に対して，キーワードフィルタしていない Tweet を入力したために十分な性能が得られなかったと考えられる．一方，追加学習モデルでは適合率 61%(83/136)であり，Baseline に比べて，抽出した正例数を増加させ，負例数を大幅に削減することができている．これは，本システムの利用により，Baseline の学習データに含まれていない新たなデータを追加学習できたためと考えられる．具体的に，追加学習モデルでは，「ハイオクガソリンに水混入　52 台中 6 台にエンジントラブル」という Tweet を新たに正例として抽出することができた．

4. まとめ

ニューラルネットワークを用いて，ニュース制作に役立つ Tweet を自動で抽出・提示するシステムを開発した．提示された Tweet をアノテートすることによって新たなデータを取得でき，この追加学習用データを用いて学習することで，システムの性能が向上することを確認できた．今後，新たに収集したデータの各カテゴリのバランスや，正例に含まれる文脈等を考慮するアルゴリズムも検討していきたい．

参考文献

[1] 足立義則，“震災ビックデータからソーシャルリスニングへ,” 放送メディア研究, No.11, pp290-293, 2014.

[2] 宮﨑太郎, 鳥海心, 武井友香, 山田一郎, 後藤淳, “Twitterからの有用情報抽出のための学習データのマルチクラス化” vol.2017-IFAT-127, no.1, pp.1-6, 2017.

[3] 萩行　正嗣，“選択式天気情報を用いたソーシャルメディアからの有用投稿抽出,” NLP2016, pp.397-400, 2016.

[4] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa, “WISDOM X, DISAANA and D-SUMM: Large-scale NLP Systems for Analyzing Textual Big Data,” COLING 2016, pp. 263-267, 2016.

[5] Seiya Tokui, Kenta Oono, Shohei, and Justin Clayton, “Chainer: a Next-Generation open source framework for deep learning,” NIPS 2015 Workshop, 2015.