E-014

事実性を考慮したニュース性のある tweet 抽出手法の検討
Examination of tweet extraction method with potential news value considering factuality

武井 友香† 宮﨑 太郎† 山田 一郎† 後藤 淳†
Yuka Takei Taro Miyazaki Ichiro Yamada Jun Goto

1. はじめに

ソーシャルネットワーキングサービス（SNS）の発展により，個人がどこでもリアルタイムに情報を発信することが可能となっている．放送局にとって，SNS 上で発信される大量の情報は，社会の出来事の把握に役立つ有力な情報源である．Twitter 等の SNS を常時チェックし，事件や事故の現場に直接居合わせた人から情報を得ることで，より迅速な報道を実現できる[1]．毎日大量に発信される tweet の中から，有益な情報を手動で取得するには多大な労力を必要とするため，我々はニュース性のある tweet を自動抽出する手法の研究に取り組んでいる[2]．実際にニュースの取材対象となった tweet を，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抽出することが可能である．しかし，自動で抽出された結果を分析すると，事件や事故を示唆する単語を含みながらも，現実の事件や事故とは無関係な tweet が抽出されていた．そこで本稿では，事実性を考慮し，ニュース性のある tweet を自動抽出する手法を提案する．

2. 事実性を考慮したニュース性のある tweet の自動抽出手法

SNS 上に投稿される文章には，事件や事故を示唆する単語を含みつつも，現実にその事象が起きていない内容の場合もある．例えば，「ちびまる子ちゃん『永沢の家火事になる（前編）』」のようにテレビ番組上での「火事」に言及した tweet である．最近では，様々なコンテンツを視聴しながら意見を共有することが流行し，テレビ番組だけでなく，ゲームやマンガなどの固有名詞を含む tweet も非常に多く存在している．また，「海外の事例を対岸の火事と楽観視できない」のように，事件や事故を示唆する単語自体が慣用句の一部となっている場合もある．これらの tweet は全て「火事」という単語を含みつつも，現実の「火事」の発生を示していない．

提案手法では，テレビ番組名，ゲーム名等，慣用句の 3 種類を特徴的な語句と定義し，特徴的な語句を含む学習データを用意する．さらにこれらの語句の有無を素性としてニューラルネットワークに加えることで，事実性を考慮し，ニュース性のある tweet を自動で抽出する．

今回は，過去にニュースの取材対象となった tweet の中で，最も大きな割合を占める火事に関する情報を自動抽出の対象とする．

2.1 ニュース性のある tweet の自動抽出

事件や事故に関する情報を含む tweet を自動で抽出するため，フィードフォワードニューラルネットワークを用いる． ニューラルネットワークへの入力は tweet の分散表現のベクトルを用いる．まず，tweet を形態素単位に分割し，Word2Vec[3]を用いて，tweet に含まれる各単語を 200 次元の分散表現に変換する．そして，このベクトルの加算平均値を tweet 全体の分散表現とする．これに，事実性を考慮する素性を付加し，学習モデルを生成する．

2.2 事実性を考慮するための素性

事実性を考慮するための基準として，テレビ番組名，ゲーム名等，慣用句の 3 種類の特徴的な語句を用いる．Wikipedia や放送局の番組表 API より，テレビ番組名 9,437 件，ゲーム名やアニメ名，マンガ名等の固有名詞 12,069 件を収集する．さらに，「火」が含まれる慣用句について，「あすとろ出版」の「故事ことわざ辞典」，「慣用句の辞典」，「四字熟語の辞典」に掲載されている 32 件を収集する．なお，この特徴的な語句リストでは，「生きる」のように一般的な動詞や形容詞になり得るタイトル，「江」のような一文字のタイトルは除外している．

表 1. 特徴的な語句の具体例

素性の種類	具体例
テレビ番組名	ひよっこ , べっぴんさん, あさイチ…
ゲーム名等	テトリス, マリオパーティ, ドラゴンクエスト…
慣用句	対岸の火事, 火事場の馬鹿力 , 電光石火…

・特徴的な語句の有無を示す入力層

tweet 本文中の特徴的な語句の有無を判定し，該当する語句が含まれる場合は，その語句の種類に対応する次元の値を「1」，含まれていない場合は「0」として入力層を 3 次元拡張する．例えば，「海外の事例を対岸の火事と楽観視できない」という投稿の場合，「対岸の火事」という慣用句が含まれるため，慣用句に対応する次元を「1」，テレビ番組名やゲーム名等についての語句は含まれていないため，それらに対応する次元の値を「0」とする．

・入力層の構成

tweet に含まれる各単語を分散表現に変換した 200 次元の入力層と，特徴的な語句の有無を示す 3 次元の入力層を区別し，それらを中間層で結合するニューラルネットワークで学習する．また比較手法として，計 203 次元を同一の入力層とするニューラルネットワークを用いる．

† NHK 放送技術研究所

3. 評価実験

提案手法の効果を確認するための 2 つの評価実験を行った．実験 1 では，特徴的な語句を含む学習データの有効性を確認する．実験 2 では，特徴的な語句の有無を示す素性の付加手法を比較する．

3.1 実験条件

学習データには正例として，2014 年 3 月から 2015 年 8 月までに，放送局の報道番組制作現場において火事に関する情報が含まれると判別された 5,065 tweet を使用した．負例は以下の 2 種類を用意する．現実の火事とは関係のない tweet のみが含まれていることを確認した．

負例 A： 2016 年中からランダムに選んだ 5,065 tweet
負例 B： ランダムに選んだ 2,533 tweet と特徴的な語句を含む 2,532 tweet を混ぜた 5,065 tweet

評価データは，1 日分の tweet 約 770 万件を，実際に放送現場で用いられているキーワードでフィルタしたものを用いる．現実の火事や火災に関する情報を含む 2,960 tweet には正例ラベル，関係のない内容の 5,686 tweet は負例ラベルを人手で付与した．

手法の実装には，Chainer [4]を利用した．入力層は手法により 200 次元または 203 次元，出力層は 2 次元とした．中間層 2 層とし，中間層のノード数は近い方から 500，250 である．また，拡張した入力層を区別する場合，入力層に対応する中間層のノード数は 2 とし，入力に近い方の層で結合させる．活性化関数 ELU（Exponential Linear Units），学習の回数は 50 回とした．Word2Vec による分散表現の生成には，2016 年 9 月の Wikipedia のダンプデータを利用し，tweet に出現する URL については除去する．

3.2 実験結果

・実験 1（学習データの比較）

ランダムに負例となる tweet を選んだ負例 A と，特徴的な語句を含む負例 B の学習データを比較した実験結果を表 2 に示す．学習データの正例は同一のデータを用いた．ここでは，素性を付加せず 200 次元の分散表現のみの学習モデルで実験する．

表 2. 学習データの比較評価結果

学習データ	Recall	Precision	F 値
負例 A(ランダム)	84.1	79.7	81.9
負例 B(特徴的な語を含む)	85.4	83.4	84.4

評価の結果，学習データに特徴的な語句を含むデータを用いると，F 値で 2.5 ポイントの性能が向上し，有効性を確認できたため，次の実験 2 では負例 B を用いる．

・実験 2（素性の付加手法の比較）

特徴的な語の有無を素性として付加し，比較した実験結果を表 3 に示す． 200 次元の分散表現のみを学習したモデルを Baseline とする．3 種類の特徴的な語句をそれぞれ 3 次元に区別したモデル，1 次元にまとめたモデルと比較する．そして，最後に 200 次元の分散表現と特徴語の有無を示す 3 次元の入力層を区別した提案手法の結果を示す．

表 3. 素性の付加による比較評価結果

Method	Recall	Precision	F 値
Baseline	85.4	83.4	84.4
+ features (3 次元)	**88.9**	82.8	**85.7**
+ features (1 次元)	82.7	83.7	83.2
提案手法	86.3	**85.0**	85.6

3 種類の特徴的な語句（テレビ番組名，ゲーム名等，慣用句）の有無を示す素性を，別々の次元として入力層を拡張することで，F 値で最大 1.3 ポイントの性能の向上が確認できた．

3.3 考察

特徴的な語句の有無を示す素性を付加せず学習モデルを生成した場合，「火事」に関連する語句が含まれている tweet が正例として抽出されていた．例えば，「火の鳥の最終回が炎上！」という投稿は，「火」や「炎上」のように火事に関連する言葉が含まれているため，正例として抽出されていた．ここで，「火の鳥」という特徴的な語句の存在を示す素性を追加することにより，負例であると判別することが可能となった．特徴的な語句の有無を示す入力層を別にすることで，正例の網羅性を維持したまま，誤抽出数を抑えることができたと考えられる．また， 比較手法において，3 種類の特徴的な語句の素性を，1 つの入力次元にまとめるよりも，3 次元に分けることで精度が向上する傾向が見られた．このように特徴的な語を区別して素性として与えることで，文中に現れる際の表記のパターンも学習できたと考えられる．

4. おわりに

本稿では，事実性を考慮しニュース性のある tweet を自動で抽出する手法を提案した．特徴的な語句を含む学習データを用い，特徴的な語句の有無を示す素性をニューラルネットワークの入力に加えることで，自動抽出の性能が向上することが確認できた．事実性を考慮することで，事件や事故とは無関係な大量の tweet の中から，ニュースの情報源となる情報により早く辿り着くことが可能となり，SNS から人手で情報取得する労力を軽減できることが期待できる．また今後，番組関連情報などからより多くの特徴的な語句を取得し，さらにリアルタイムに取り入れていくことで更なる性能の向上を目指したい．

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, “ニュースに制作に役立つ tweet の自動抽出手法,” NLP2017, pp.418-421, 2017.
[3] Tomas Mikolov, Kai Chen, Greg Corradoet, and Jeffrey Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.
[4] Seiya Tokui, Kenta Oono, Shohei, and Justin Clayton, “Chainer: a Next-Generation open source framework for deep learning,” NIPS Workshop, 2015.