インフルエンザ患者数推定のためのツイート分類手法

遠藤圭太† 苅米志帆乃‡

†‡長野工業高等専門学校 電気電子工学科 〒381-0041 長野県長野市徳間 716

E-mail: †14207@g.nagano-nct.ac.jp, ‡s_karikome@nagano-nct.ac.jp

あらまし インフルエンザは全ての年齢層に対して感染し，世界中で繰り返し流行している．流行を事前に予測することにより早期に対策でき，予防につながる．流行を知る一つの手段として，Twitterがある．Twitterは代表的なSNSの一つであり，情報をいち早く入手できるという特徴がある．本研究では，インフルエンザの患者数推定をするために，インフルエンザに関するツイートを4つに分類し，推定をする手法を提案する．有効性を評価するために，実際のツイートを収集し，評価実験を行った．

キーワード Twitter，インフルエンザ，情報抽出，情報分類

1. はじめに

インフルエンザは全世界で毎年300～500万人が重症化し，29～65万人の死者を出している．主な症状は急な発熱，咳，頭痛，関節痛などで風邪の症状に似ている．一年を通して世界中で流行しているが，日本では冬に最も流行する[1]．また，近年のスマートフォンの普及とともに，SNSもユーザ数が増加してきた．

Twitterは「ツイート」と呼ばれる短文を投稿するSNSである．「天気がいい」「今日も頑張ろう」など，ユーザが感じたこと，身の回りに起こったことなどを手軽に投稿できる．体調が悪い時に「頭が痛い」「熱がある」といったツイートをするユーザもおり，インフルエンザ患者の一部は「インフル3日目」，「インフルB型なう」などといったツイートを投稿する．他のSNSと比べてTwitterが優れている点は，情報伝達の早さである．また，インフルエンザは冬に非常に早いペースで流行するため，流行の予測をするには常に最新の情報が必要になる．そこでTwitterのツイートを活用すると，ある程度のインフルエンザの流行予測をすることができる．荒牧ら[2]はツイート数と現実の統計量の間の時間的ギャップや空間的ギャップの補正をするために，インフルエンザ患者数推定を事例にして提案手法を検証した．遠距離言及発言はネット上での注目度を示しており，遠距離言及発言とインフルエンザ患者数は一定の相関があることを実証した．遠距離言及発言により注目を浴びるにつれ，話題としての価値が下がり，関連する発言が減少するというソーシャルセンサの劣化モデルを提案した．このモデルを使うとある地域における特定の日のインフルエンザ患者数を求めることができる．本研究では，荒牧らが作成したソーシャルセンサの劣化モデルを使用するための分類手法を提案する．モデルを使いインフルエンザ患者数を算出するにはインフルエンザに関するツイート数が必要になるが，その関連発言をいくつかのパターンに分けることで，モデルよりもさらに早い段階で推定が行えるようになると考えた．本研究は，Twitterのツイートからインフルエンザに関するツイートを抽出，分類し，患者数推定をするためのツイートの分類手法を提案することを目的とする．ツイートの分類パターンを表1に示す．

表1 分類パターン

	ツイートの種類
(a)	ニュースやリツイートに関するツイート
(b)	インフルエンザの予防接種に関するツイート
(c)	インフルエンザの疑いがあるツイート
(d)	インフルエンザであるツイート

先行研究では(a)と(d)を収集していたが，本研究ではこれに加え，(b)と(c)を入れた4パターンに分類する．

(a)は，荒牧らの関連研究で提案されたモデルの式にニュースやリツイートに関するツイートが必要であるため分類する．(a)に分類されるツイートはネットニュース記事のリンクが貼られているツイートや，「北海道でインフルが流行ってるってニュースやってた」などといった，ニュースに関するツイートが分類される．(b)は(a)と同等の性質があると考え，分類する．「予防接種してきた」といったツイートが分類される．(c)には「熱も高くて喉も痛いしインフルかも」といった，確実ではないがインフルエンザと疑われるツイートが分類される．(d)には「インフルA型だった」といった，確実にインフルエンザであるツイートが分類される．このような手法にすることで，ツイートを自動で分類できるという利点がある．

さらに荒牧ら[3]はTwitterのインフルエンザに関する発言を，機械学習手法でツイートを分析する手法を提案した．インフルエンザにかかった人物が存在するかを判断し，インフルエンザ流行予測をした．この関連研究と本研究との違いは，関連研究ではインフルエンザ陰性，陽性という2値判定をしているが，本研究

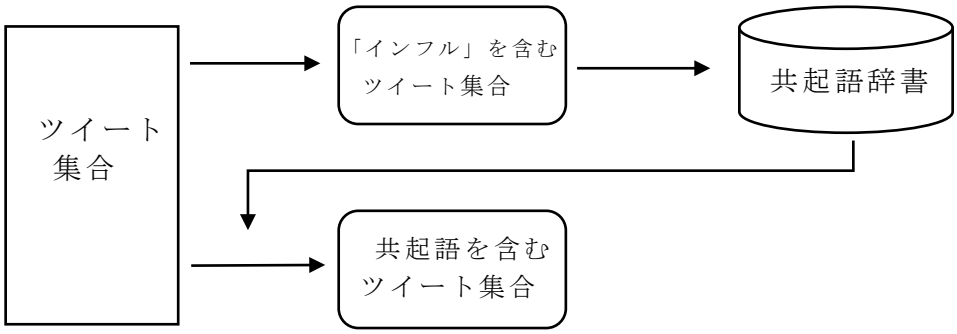

図 1　ツイート収集の概要

はいくつかのパターンに分類するという点である．これにより，少ないツイート収集量で予測ができるという利点がある．

梅島ら[4]は Twitter におけるデマとデマ訂正の拡散の傾向を分析し，災害時のデマ情報を自動取得する手法の提案を行った．Twitter のリツイートをいくつかのパターンに分類し，リツイートされやすいツイートを仮説，実験により考察した．

2. ツイート分類手法

2.1 概要

はじめに，ツイートを収集と共起語辞書の構築を行い，その次にツイートを 4 つに分類する．2.2 節でツイートの収集と共起語辞書の構築について述べ，2.3 節でツイートの分類について述べる．

2.2 ツイート収集および共起語辞書構築

図 1 にツイート収集の概要を示す．3 つの手順で収集を行う．

(1) 「インフル」を含むツイートの収集

インフルエンザに関するツイートを収集するために，「インフルエンザ」とその略称である「インフル」のどちらかが含まれるツイートを収集する．

(2) 共起語辞書の構築

(1)以外のインフルエンザに関するツイートを収集するために，本研究では共起語を用いる．(1)で収集したツイート集合をもとに，インフルエンザに関するツイートに頻出する名詞および複合名詞を収集し，共起語辞書を構築する．しかし，インフルエンザにあまり関係がない語も抽出されてしまうため,「インフル」が含まれていない一般のツイートも収集し，そこから得られた共起語から共通する語を除外する．すると,「インフル」が含まれるツイートのみ頻出する語をある程度絞ることができる．

(3) 共起語を含むツイートの収集

共起語辞書の語がツイート中に 1 つ以上含まれているツイートを収集する．そこで(2)で構築した辞書に登録されている語を含むツイートを収集することでインフルエンザに関するツイートを収集する．例として,「高熱」「関節痛」という名詞が共起語辞書に含まれている場合,「高熱で関節痛もひどい」というツイートを収集することができる．

2.3 ツイートの分類

2.2 節の(1)および(3)で収集した 2 つのツイート集合を 4 パターンに分類する．

具体的な分類方法は，(a)～(d)のそれぞれで分類ワードを指定し，ツイートに分類ワードが含まれていたら(a)～(d)のいずれかに分類をする．分類ワードの一覧を表 2 に示す．

表 2　分類ワードの一覧

	分類ワード
(a)	ニュース，http
(b)	予防接種，注射
(c)	熱，鼻水，頭痛，寒気，喉，検査，感染，病院，学級閉鎖，欠席
(d)	陽性，インフル A，インフル B，A 型，B 型

例として，ツイートに「予防接種」というワードが含まれていたら(b)に分類される．特に(c)については，インフルエンザが風邪の症状に似ていることから，風邪の症状や,「感染」「学級閉鎖」などインフルエンザの特徴語を指定した．分類の順番は(a)，(b)，(c)，(d)，(e)の順番なので,「頭痛と寒気がひどい．インフル陽性かもしれない.」というツイートは(c)に分類される．

3. 評価実験

3.1 概要

ツイートの分類をするための提案手法の評価実験を行った．

3.2 共起語辞書の構築

まず，ツイートを収集するための共起語辞書の作成を行った．対象期間は 2018 年 1 月で,「インフル」が含まれているツイートと,「インフル」が含まれていない一般ツイートのそれぞれについて，各日 50~55 件収

集した．表 3 に「インフル」が含まれているツイートの共起語辞書と一般ツイートの共起語辞書，表 4 に共通の語を除外した共起語辞書を示す．表 3 の太字になっている語は，2 つの共起語辞書に共通して出現している語である．

表 3　2 つの共起語辞書

出現頻度順位	インフル	一般
1	**ん**	**の**
2	**の**	**ん**
3	熱	**今日**
4	**今日**	**ー**
5	**ー**	**こと**
6	風邪	**人**
7	病院	**私**
8	**人**	**今**
9	**こと**	**)**
10	**昨日**	**(**
11	**私**	**方**
12	**)**	**よう**
13	検査	9 時
14	**仕事**	**さん**
15	**よう**	**自分**
16	**気**	追加
17	**そう**	リツイート
18	**今**	**仕事**
19	**(**	貴女
20	**これ**	もの

表 4　共通語を除外した後の共起語辞書

出現頻度順位	インフル
1	熱
2	風邪
3	病院
4	検査
5	マスク
6	休み
7	職場
8	会社
9	喉
10	体調

共通語を除外することで「インフル」が含まれるツイートにのみ頻出する共起語に絞った.その結果,「熱」「風邪」「病院」など，インフルエンザに関係の深い語を共起語辞書にすることができた．さらに,「ん」「の」などの口語や記号など，形態素解析で誤って名詞と判断された語も除外することができた．しかし，共通語でなければ除外できないので，今後は「サ変接続」「接尾辞」など，名詞のカテゴリを細かいところまで判断し，一般名詞でないものをあらかじめ除外しておくという手法が必要と考えられる．

3.3　ツイートの収集

次に,「インフル」が含まれるツイート 310 件と共起語を含むツイート 310 件を収集した．対象期間は 2017 年 12 月で，各日 10 件ずつ収集した．共起語を含むツイートについては，表 3 に示した共起語辞書の上位 10 件を使用し，ツイートにこの 10 件の語が 1 つ以上含まれているツイートを対象に収集した．

3.4　評価尺度

4.3 節で収集したツイートをまずは人手で分類し，それを正解データとした．次に自動で分類し，その再現率と精度で評価した．

$$再現率=\frac{(抽出されたツイート中の正解ツイート数)}{(正解ツイート数)}$$

$$精度=\frac{(抽出されたツイート中の正解ツイート数)}{(抽出されたツイート数)}$$

3.5　実験結果

「インフル」が含まれるツイートの分類結果を表 5，共起語が含まれるツイートの分類結果を表 6 に示す．

表 5　「インフル」が含まれるツイートの分類結果

	再現率		精度	
(a)	8/12	0.667	8/15	0.533
(b)	32/37	0.865	32/51	0.627
(c)	34/71	0.479	34/74	0.459
(d)	9/74	0.122	9/12	0.750
(a)～(d)以外	69/116	0.595	69/158	0.437

表 6　共起語が含まれるツイートの分類結果

	再現率		精度	
(a)	−	−	0/18	−
(b)	−	−	−	−
(c)	13/27	0.481	13/74	0.176
(d)	−	−	−	−
(a)～(d)以外	190/283	0.671	190/218	0.872

表 5 について，(b)の値は 0.865 と高く，(d)の値は 0.122 と低くなった．精度は再現率ほどの数値のばらつきはなかった．(d)は再現率が低かったのに対し，精度は全体の中で一番高い数値となった．

表 6 について，傍線部は該当するツイートが無かったことを示しており，収集した 310 件の中に(a)，(b)，

(d)に該当するものがなかった．(c)と(a)~(d)以外の再現率は値の差があまりなかったが，精度は差が大きくなった．(c)の精度は 0.176 と低い値だったが，(a)~(d)以外の精度は 0.872 と高い値となった．

表 5 と表 6 を比較すると，再現率に大きな差は見られなかったが，精度は(c)および(a)~(d)以外ともに値が大きく違っていた．

表 5 で(a)~(d)以外に分類されたツイートは「インフル嘘だった」や，「インフルと勘違いされた」といった内容のツイートだった．表 6 で(a)~(d)以外に分類されたツイートは「ぼくも子どもの体調見ながらいろいろやってるし…」や，「首痛すぎて会社行きたくない」といった内容のツイートだった．

3.6 考察

表 5 について，(b)の値は高く，(d)の値は低くなった要因は，(b)は予防接種についてのツイートで言葉が限定的であるのに対し，(d)は「インフルだった」「インフル確定」など様々な表現が使われるためであると考えられる． (d)は再現率が低かったのに対し，精度は全体の中で一番高い数値となった要因は，(d)を分類するために指定した語が，確実にインフルエンザだと思われる場合に使われることがほとんどであるためだと考えられる．

表 6 について，(a)~(d)以外の精度が高くなった要因は，共起語が含まれるツイート集合にインフルエンザに関するツイートが少なく，精度が高くなったことが考えられる．

表 5 と表 6 を比較したとき，精度が(c)，(a)~(d)以外ともに値が大きく違っていた原因として，表 5 の(c)は，「インフル」が含まれているツイートしかないため，ある程度の精度が出たことが考えられる．

(a)~(d)については，分類したのち，モデルの式に代入するために重みづけをするが，(d)はインフルエンザの流行に直接関係しているツイートのため，(d)に分類されたツイートの重みを一番大きくし，(c)， (b)，(a)という順番で小さくしていくと良いと考えられる．また，共起語が含まれるツイートでは(a)，(b)，(d)に該当するツイートが収集できず，十分な考察ができないため，収集件数をさらに増やす必要がある．

4. おわりに

本研究はインフルエンザの患者数を推定するためのツイートの分類手法を提案し，評価を行った．ツイートの収集では，「インフル」を含むツイートの他に，共起語を含むツイートも収集した．これにより，「インフル」が含まれていなくてもインフルエンザの疑いのあるツイートを収集することができた．ツイートの分類では，モデルにおける分類パターンを新たに 2 つ追加し，計 4 パターンで分類を行った．実験結果を見ると，再現率や精度でまだ改善の余地がある．これらを改善するためにはツイートを分類するために指定した語を再検討する必要があると考えられる．

参 考 文 献

[1] World Health Organization . Influenza (Seasonal). https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal), (参照 : 2018 年 12 月 29 日).

[2] 荒牧英治，若宮翔子．ツイート数と現実の統計量との差異に関する検討．統計数理第 64 巻第 2 号，Proceedings of the Institute of Statistical Mathematics Vol.64, No.2, 233-246，2016.

[3] 荒牧英治，増川佐知子，森田瑞樹．Twitter Catches the Flu: 事実性判定を用いたインフルエンザ流行予測．IPSJ SIG Technical Report，Vol.2011-NL-201 No.1. ，Vol.2011-SLP-86 No.1，2011.

[4] 梅島彩奈，宮部真衣，荒牧英治，灘本明代．災害時 Twitter におけるデマとデマ訂正 RT の傾向．IPSJ SIG Technical Report，Vol.2011-DBS-152 No.4，Vol.2011-IFAT-103 No.4，2011.