統計数理(2016)
第 64 巻 第 2 号 233–246

特集「統計的言語研究の現在」
［研究ノート］

ツイート数と現実の統計量との差異に関する検討

荒牧 英治†・若宮 翔子†

(受付 2015 年 12 月 31 日；改訂 2016 年 11 月 10 日；採択 11 月 21 日)

要　　旨

ソーシャルメディアサービスの普及により，人々や社会の状況を調査する新たなアプローチが開拓された．この結果，インフルエンザや地震などを対象とした多くのサーベイランスや監視システムが提案され，現在も稼働している．しかし，ソーシャルメディア上のユーザ発信データ(発言内容，時間や場所)が必ずしも現実を正確に反映しているとは限らない．例えば，デマや流言などが出現することもあり，新聞などの既存のメディアと比べて，内容の信頼性は十分ではなく，時間的または空間的な正確性にも限界がある．本稿では，ソーシャルメディアを代表する Twitter を用いて構築したインフルエンザ・サーベイランス・システムを例に，ツイート数と現実の統計量の時間的なずれと空間的なずれについて検討し，背後にあるバイアスについて議論する．

キーワード：ソーシャルメディア，Twitter，自然言語処理，ソーシャル・コンピューティング，インフルエンザ．

1. はじめに

近年の情報処理の発展は World Wide Web(以降，Web)の存在なしに語ることはできない．Web はかつてないほどの巨大なデータを含み，かつ，誰もが発信できるメディアである．この特性を活かし，Web ならではの新たな研究分野も形成されている．評判情報抽出 Pang et al. (2002) がその代表例である．さらに，近年では，Twitter や Facebook などのソーシャルメディアが爆発的に普及し，ここから，評判だけでなく，より詳細な情報を抽出する試みにも関心が集まっている．ソーシャルメディアのデータは，大規模かつ即時的であり，一部には位置情報も付与されているなど，これまでの Web データには見られなかった特徴がある．この特徴を利用し，疾病サーベイランス(Aramaki et al., 2011; Paul and Dredze, 2011; 谷田 他, 2011)，地震検知(Sakaki et al., 2010)，選挙結果予測(Tumasjan et al., 2010)や株価予測(Bollen et al., 2011)など，その応用領域は多岐にわたっている．

これらの研究は，暗に Web テキストが現実に対応しているという仮定がベースとなっている．しかし，Web テキストは必ずしも現実と正確に対応しているわけではなく，様々なバイアスから両者の間にギャップが生じる場合がある．最も大きなバイアスの一つは，SNS ユーザの偏りによるものである．例えば，人口 1 万人あたりの Twitter ユーザは，東京都では 369.82 人であるのに対し，最も少ない佐賀県では 63.67 人であり，約 6 倍もの差がある(odomon.net, 2013)．さらに，18 歳から 24 歳のユーザが多い．これらを考えると，都市部の若者を中心にデータを採取していることになる．これ以外にも，問題となりうるバイアスは多く存在し，現

† 奈良先端科学技術大学院大学：〒630–0192 奈良県生駒市高山町 8916–5

実とソーシャルメディア・データとの定量的な差異を生み出している．

本研究では，ソーシャルデータを実世界の現象の「センサ」として用いる先行研究に対し，現実とソーシャルデータに存在する時間的，空間的なギャップを補正し，より高い精度で実世界の現象をモデル化することを目指す．本稿では，感染症サーベイランスと呼ばれる感染症流行の把握のために，Twitter のようなユーザ投稿発言データの利活用が進みつつある（国立研究開発法人日本医療研究開発機構（AMED），2015）という背景を受け，主な感染症の一つであるインフルエンザを対象に，代表的なソーシャルメディアである Twitter を用いて構築したサーベイランスシステムを実例として，時間的なギャップ（2 章）と空間的なギャップ（3 章）について議論し，これらのギャップを補正した方法をインフルエンザの患者推定という事例（4 章）を挙げて考察する．5 章で関連研究を紹介し，6 章でまとめを述べる．

2. 時間的ギャップ

本章では，ソーシャルメディアと現実がどのような時間的なギャップを持ちうるのか，時系列データであるインフルエンザの流行を題材に議論を行う．

2.1 材料: インフル・コーパス

インフルエンザに関する Twitter 上での発言を集めたコーパス（以下，インフル・コーパス）を用いて，時間的なずれの調査を行った．インフル・コーパスはインフルエンザ・サーベイランスサイト（奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室, 2016）を稼働し，収集されたデータをもとにしたコーパスであり，以下の手順で構築されている．

まず，2008 年 11 月から 2010 年 7 月にかけて Twitter API を用いて 30 億発言を収集し，そこから「インフル」を含む発言（インフル関連発言）を 10,443 件無作為に抽出した．これに対して作業者が，発言者がインフルエンザ罹患者（正例）であるか否か（負例）という事実性を判定し，ラベル付けした．アノテーションについては，以下のような基準に照らし，一つでも該当するものがあれば，負例とみなした．より詳細な基準に関しては，アノテーション・ガイドライン（Aramaki and Wakamiya, 2016）を参照していただきたい．

(1) 発言者または発言者と距離的に近い人物（同一都道府県近郊の人間）の疾患でない場合
(2) 現在または近い過去（24 時間以内）の疾患でない場合
(3) 否定，疑問や「かもしれない」といった事実でない場合

このトレーニング・データをもとに，2011 年 8 月からインフルエンザ・サーベイランスサイトを運用し，現在までに，8,129,571 件のインフル関連発言を抽出し，インフルエンザ罹患者（正例）による発言であるか否か（負例）という事実性に関するラベルを自動推定した．なお，全ての単語の表層系を素性とし，SVM を用いて構築した分類器を用いた．

この結果，テストデータの 58% が正例と判定された．F 値を求めたところ，0.76 という高い精度を示した（Aramaki et al., 2011）．

2.2 結果

2012 年から以降 3 年間のシーズン時の発言データから，自動判別で得られた正例（以降，単に正例と表す）の発言データを抽出し，インフルエンザの患者数を下記の式により推定した．

$$I_0(a,t) = \bar{I_0} \cdot \frac{M(a,t)}{N(a)} \tag{2.1}$$

ここで，$M(a,t)$ は対象地域 a における特定の日 t のインフルエンザの正例数，$N(a)$ は対象地

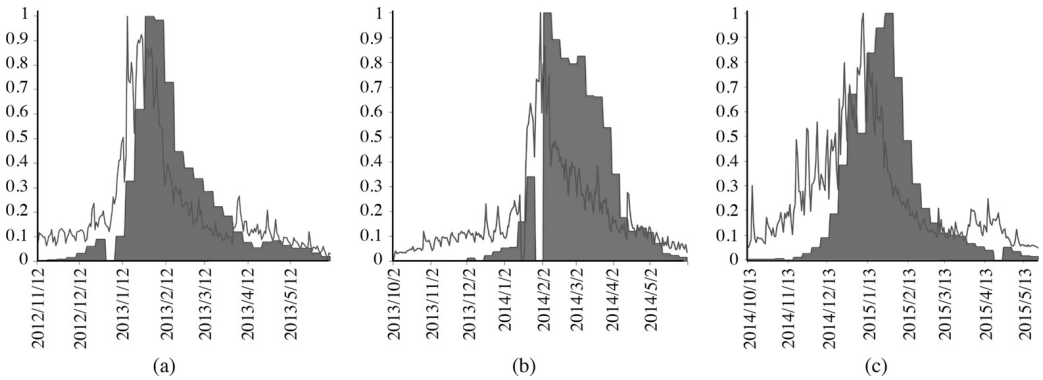

図 1. 2012–2014 年度のインフルエンザの患者数と正例の発言を用いて推定した患者数．横軸は日付 d，縦軸は患者数(塗り潰し)と Twitter 発言数(折れ線)とする．それぞれの値は，年度の最大値を 1 として，0 から 1 の値になるように正規化して表示している．

域 a におけるソーシャルセンサ数(ユーザ数)，$\bar{I}$ はスケールパラメータである．なお，本研究では，発言数はユーザ数に比例するものとみなしている．そのため，ソーシャルセンサ数は，各地域における対象期間中の平均発言数に基づき算出される．結果を図 1 に示す．2011 年については，Twitter API の仕様変更に伴い，一時期クロールが停止していたため，本研究では対象外とした．

正解データのインフルエンザの患者数として，国立感染症研究所 感染症情報センターが報告している患者数を用いた．これは，国立感染症研究所のホームページにおける患者発生状況の「都道府県別報告数・定点当たり報告数」より取得可能であり，秋から春のインフルエンザシーズンにかけて，都道府県ごとのデータが毎週公開される．発言数より推定された患者数と実際の患者数との差分を誤差として調査を行った．誤差を算出する際，年度ごとの最大値が 1 となるようにそれぞれの値を正規化している．誤差は正規化した推定患者数の値から正規化した実際の患者数の値を引いた値であり，そのスケールは −1 から 1 である．正の誤差は，推定患者数が実際の患者数を上回っていることを意味し，反対に負の誤差は，推定患者数が実際の患者数を下回っていることを意味する．なお，この誤差の時間ごとの和が小さくなるように，発言数から患者数を推定するときのパラメータ $\bar{I}_0$ を調整した．

一般的に，感染症の把握／予測は，推定した値と患者数との相関係数が評価基準であり，ここでいう誤差の最小化は，間接的にこの評価をよくすることができる．

いずれの年においても，実際の流行のピーク前に，Web 上でのインフルエンザ発言が増加し，逆に，ピーク後は実際の流行度合いよりも発言が減少する，という傾向が確認された．対象とした 3 年間の平均をとった結果を図 2(a)に示す．

全体を平均すると，3 年間の平均誤差は −0.0178(標準偏差 0.188)となり，実際の患者数がソーシャルメディアの発言数を上回っているが，ピーク前，ピーク直後，平常時と 3 つの異なる状態があることが分かる．

まず，ピーク前は平均誤差が 0.136(標準偏差 0.062)となり，ソーシャルメディア上での発言数が実際の患者数を上回り続ける．標準偏差が小さいことから，この誤差は安定しており，ソーシャルメディア上では現実より常に加熱した状態となっているといえる(図 2(b))．

次にピーク後は，平均誤差が −0.256(標準偏差 0.122)となり，今度は逆に，実際の患者数がソーシャルメディアの発言数を上回っている．標準偏差が大きいことから分かるように，この

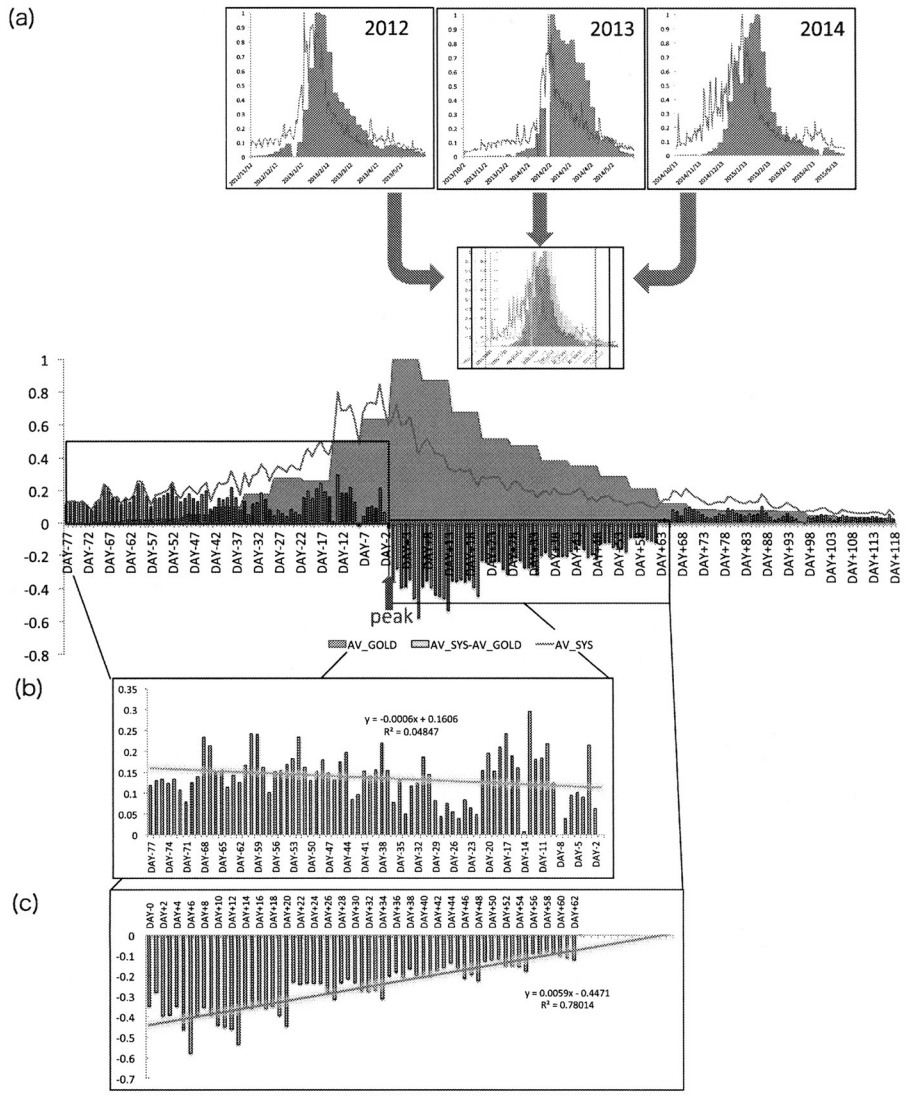

図 2. 各年度の患者数がピークとなる日時を合わせた平均．横軸は患者数のピーク日時からの相対的な日付，縦軸は，以下で定義される患者数の平均(塗り潰し)，Twitter 発言数の平均(折れ線)，および両者の平均誤差(棒)を示す．

上回り方は時期によって異なる．ピーク直後は最も誤差が大きく，0.40 に近い．この誤差は次第に減少し，最後には 0.10 付近となる(図 2(c))．

最後に，平常時は，ピーク前と同様に，ソーシャルメディア上での発言数が患者数を上回る状態が安定して続く．

2.3 考察

この結果から，次の 2 つの知見が得られた．まず，誤差は常に存在しており，(1)ピーク前，(2)ピーク直後，(3)平常時の 3 つの段階がある．(1)ピーク前と(3)平常時における誤差は似て

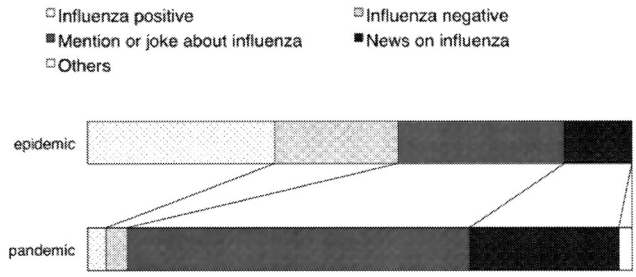

図 3. インフルエンザに関連する発言の分類．“Influenza positive” はインフルエンザ患者の発言．“Influenza negative” は「インフルエンザでなくてよかった」など否定を含む発言．epidemic（2011 年 11 月）および pandemic（2009 年 1 月）より，それぞれ 200 発言を無作為抽出し，人手で分類した．

おり，発言数が実際の患者数を上回る．一方，(2)ピーク直後では，発言数よりも患者数が多くなり，徐々に誤差が減少し，(3)平常時へと収束する．この性質を利用すれば，ピークを知ることにより，流行推定の精度を向上させることができる．

次に，ソーシャルメディア上での発言数のピークがいつであるかを事前に知ることは困難である．(1)ピーク前はほぼ単調な誤差が続くばかりであり（図 2(b)），いつピークを迎えるのかを伺わせる手がかりはない．つまり，予測を行うことは難しい．

この 2 つの知見，すなわち (1)誤差はピーク前には実際よりも発言が多く，ピーク直後には発言が少なくなること，(2)ピークがいつであるかを知ることは困難であること，はインフルエンザ調査の実用的応用を困難にしている．

このような発言量の減少が生じる原因について考察する．図 3 は，その一端を示す例であり，平常時（epidemic, 2011 年 11 月）と WHO が新型インフルエンザへの懸念を表明した時期（pandemic, 2009 年 1 月）における発言の内訳を示している．epidemic 期には，「インフル」を話題にした発言やニュースに対する発言が増加する．高まる不安や対応法への懸念として「インフル」という単語が使用されるといえる．このような epidemic 期を過ぎると，すでに多くのユーザは，大量のニュースで「インフル」を目にしていることになる．このようにインフルエンザ流行がすでに周知され一般に既知となった状態では，自分（または周囲）にインフルエンザが発生したとしても，話題としての価値が低下しているため，ソーシャルメディア上で発言することを躊躇すると考えられる．これは，図 1 の Twitter 発言数の推移にも表れており，患者数がピークを迎えると発言数は大きく減少しており，情報伝搬のタイミングがギャップを生み出している可能性がある．

3. 空間的ギャップ

本章では，ソーシャルメディアと現実がどのような空間的なギャップを持ちうるのか，位置情報が付与されており，かつ，ランドマーク表現が含まれる発言を用いて調査する．なお，ランドマーク表現が含まれる場合には，次の可能性がある．

(1) ランドマークが地名を示す
- ランドマーク上にいる
- ランドマーク上にいない

(2) 地名を示さない

先行研究においては，(1)と(2)を自動分類するアプローチ(Awamura et al., 2015) もあるが，本研究では(Antoine et al., 2015)と同様に，これらを区別せずに，どのような場所が言及されるか調査し，議論する．

3.1 材料: 京都 GPS コーパス

空間的なずれを調査するために，京都を対象地域として，京都市近郊で発信された GPS 情報付き発言(以下，**京都 GPS コーパス**)を用いた．京都 GPS コーパスは，Twitter API を用いて約 1 年間(2011 年 7 月 15 日から 2012 年 7 月 31 日)にわたり収集した約 3.7 万件発言から構成されている．

次に京都市近郊の場所として，以下の 4 つのタイプの 6 つのランドマークを選択した．(1)広域領域内に複数の施設を持つ広域複合ランドマーク((a)同志社大学，(b)京都大学)，(2)局所的な特定の施設からなる狭域単一ランドマーク((c)河原町駅，(d)四条駅)，(3)広域領域だが単独の施設からなる広域単一ランドマーク((e)京都府立植物園)，(4)境界を持たないランドマーク((f)吉田)．図 4 に，これら 4 つのタイプを代表する 6 つのランドマークとそれらを含んだ発言をそれぞれの位置情報に基づき地図上にマッピングした結果を示す．

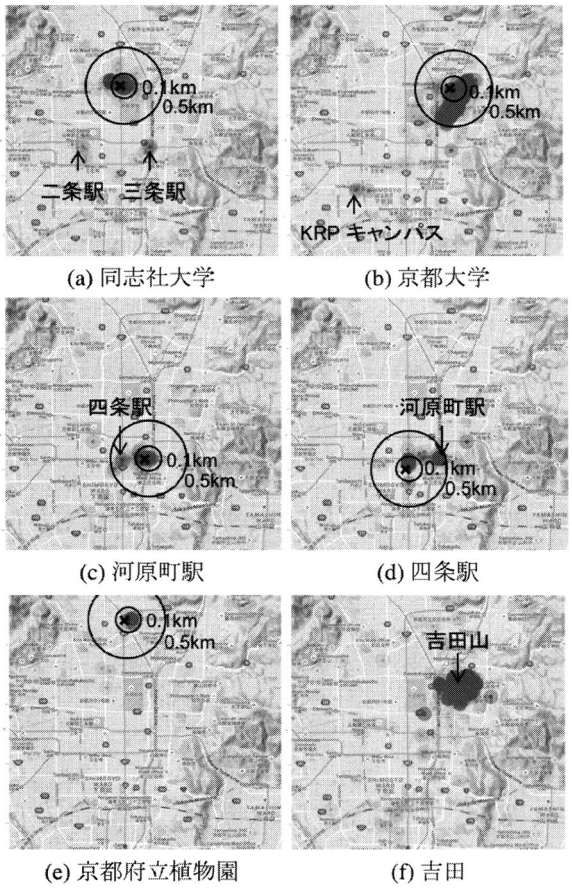

図 4. 6 つのランドマークに関する発言の地理的分布．× 印はランドマークの場所であり，2 つの円はそれぞれランドマークの場所から半径 0.1 km, 0.5 km の領域を示す．

3.2 結果

図 4 より，多くの場合，特定のランドマークに関する発言はそのランドマークの場所から 0.1 km 以内に集中しており，自分が滞在している場所について，言及していることを伺わせる．図 5 は，京都大学，四条駅，京都府立植物園に関する発言の距離ごとの密度である．多少異なった場所においても小さなピークが出現していることが分かる．

タイプ別に見れば，大学については，(a)「同志社大学」は，広域領域内に複数の施設を持つが，そのほとんどの発言は，「同志社」周辺に集中し，顕著なずれは見られない(図 4(a))．ただし，二条駅や三条駅といった周辺の駅にもピークが見られる．一方，同じく広域領域内に複数の施設を持つ(b)「京都大学」に関しては，図 4(b)のように京都大学の別キャンパスである KRP(京都リサーチパーク)キャンパスや最寄り駅である出町柳駅周辺での発言が多く見られる．このように，これからその場所へ向かう，または，離れる際に発言が行われ，実際の位置よりずれる現象が見られる．

次に駅を見てみる．(c)「河原町駅」は，この場所周辺での発言数が相対的に多いものの，四条駅付近についても発言が多い(図 4(c))．同様に，(d)「四条駅」に関しても，河原町駅周辺でも多くの発言がなされている．これは，河原町駅にいる人々が京都駅に向かう場合，四条駅を経由する行き方が一般的であるため，これから移動する場所，または，直前までいた場所について言及するという傾向が反映された結果であると考えられる．

最も集中して発言が見られたのは，上記の中でも京都大学に次ぐ大きな面積を持つ(e)「京都府立植物園」である(図 4(e))．広域であっても，発言箇所が入園時の入り口付近に集中している．

最後に (4)境界を持たないランドマーク(f)「吉田」の結果を示す．このランドマーク名は，「吉田(神社)」「吉田(山)」「吉田(寮)」など周辺の複数の地名を表し多義的であり，かつ，明確な境界を持たない語である．このため，「吉田山」を中心として発言がなされているものの，広域に発言が分散している(図 4(f))．

3.3 考察

特定の場所に関する場所参照発言がどこからなされているのかというパターンは場所ごとに依存するが，最寄り駅(同志社大学や京都大学)や隣接駅(河原町駅や四条駅)，出入り口付近(植物園)など，そのずれには一定の傾向があることが多い．これを，あらかじめ知ることができれば，GPS ベースの位置情報が付加されていない発言であっても，発言内容をもとに位置情報を推定することが可能になる．例えば，広域であっても施設が単一あるいは少数であれば，京都府立植物園のように出入り口付近にいると推定可能な場合がある．逆に，遠くの場所から言及される場合は，これから移動する予定，または，移動経路について言及しているなどのパ

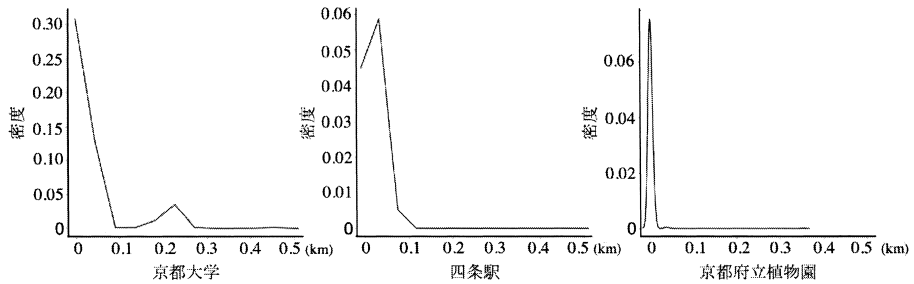

図 5. 3 つのランドマークに関する発言の距離ごとの密度.

ターンが見られる場合がある．この場合，どのような理由で遠くから言及するかを推察することで，より高い推定が可能になると考えられる．このようなずれは，発言を前後の文脈を考慮して自然言語処理により解析することにより，分離できる可能性がある．

4. 時間的・空間的なギャップの補正：インフルエンザ患者数推定を事例として

ここまで時間的なギャップ(2 章)と空間的なギャップ(3 章)について述べてきた．2 章では，ピークを過ぎると発言量が低下すること，その原因としては，ピーク後の話題としての価値低下が考えられることについて述べた．3 章では，空間的なギャップの原因にはある種の定型性(例えば，移動経路や最寄り駅)があることを示した．本章では，これらを考慮しつつ，インフルエンザ推定精度の向上を試みる．

図 6 に典型的なインフルエンザ流行の推移(2013 年北海道)を示す．感染症情報センターの値とソーシャルメディアから得られた 2 つの値が描かれている．ソーシャルメディアに基づく値は，北海道内でのインフルエンザに関する発言と，空間的なギャップを含む発言(以降，**遠距離言及発言**)により算出される．ここでいう遠距離言及発言とは，北海道以外の地域から北海道のインフルエンザについて言及している(ここでは，単語「北海道」を含む)発言とする．

前者は，ピーク後に実際の患者数に相対して大幅に減少し，2 章で述べた通りの時間的なギャップを示している．後者(遠距離言及発言)は，数は少ないものの，数回程度のバースト(急峻な盛り上がり)が存在し，特にピーク時に大きな盛り上がりを見せている．また，ピーク後には大きなバーストはない．

ピーク前にバーストが存在する理由について考察する．遠距離言及発言は，自分がいない離れた場所でのインフルエンザに関して言及しており，実際は，(1)ユーザ本人が移動した結果，遠隔地への発言となる，または，(2)ニュース・メディアなどの二次情報を通して遠隔地のインフルエンザ流行を知り，それについて発言する場合が考えられる．(1)の例として，「新型インフルエンザかぁ…北海道のライヴに遠征にいって家に帰ったら発症したんだよなぁ…絶対あのライヴ会場に発症者居たよなｗ」(千葉県より発信)，「北海道にてインフルエンザになった。

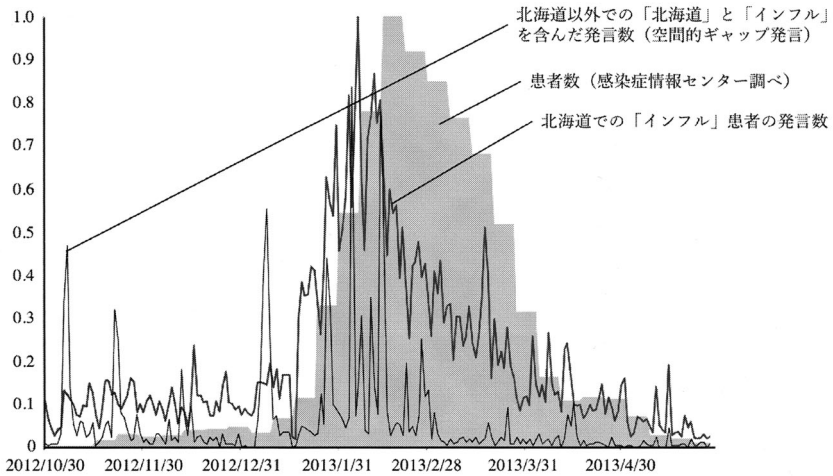

図 6. X 軸は日付．Y 軸は，2013 年北海道におけるインフルエンザ関連の発言量，遠距離言及発言，感染症情報センターの報告を示す．なお，シーズン内の最大値を 1.0 として各値を正規化している．

まじ、ありえないわ」(東京より発信)といった発言があった．しかし，大勢の人間が同時期に移動することは考えにくいため，(1)は通常はバーストの原因となりにくい．

すなわち，インフルエンザ流行のピーク前，またはピーク中に，ニュース・メディアで取り上げられたことをきっかけに，(2)のニュースの引用(例えば，「【■■ニュース】北海道で例年より早くインフルエンザ大流行」)や，流行場所を懸念する発言(例えば，「@■■ いわく北海道でインフルエンザ蔓延中」)などが大量に発生すると考えられ，バーストを形成することになる．なお，北海道以外から北海道のインフルエンザに関する発言を 412 件取得し，ニュース・メディアなどの二次情報を通して知り，発言していると思われる発言を人手で調査した．その結果，(1)ユーザ本人が移動した結果，遠隔地への発言は約 10%，(2)ニュース・メディア，あるいは，北海道に住む家族や知人などからの情報に対して発言は約 23% であった．残りの発言の大半は，北海道とインフルエンザについてそれぞれ別の文脈で述べているものや，北海道のインフルエンザについて述べているが負例のものであった．さらに，北海道におけるインフルエンザ流行のピーク中(2014 年 2 月 15 日)に，北海道以外から北海道のインフルエンザに関して発信された 100 件の発言を調査した．その結果，約 9 割の発言が(2)ニュース・メディア，あるいは，北海道に住む家族や知人などからの情報に対しての発言であった．

このように，遠距離言及発言がネット上でのある種の注目度を示しているとすると，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モデルを考えることができる．

4.1 ソーシャルセンサの劣化モデル

遠距離言及発言によって，インフルエンザ関連発言が減少していく過程を以下のようにモデル化する．

ピーク前：インフルエンザ流行前にインフルエンザに罹患したユーザは，インフルエンザ関連発言を行う．

ここで，N 人の Twitter ユーザによる発言が T 個あるとき，発言数 T はユーザ数 N にそのまま比例するものとみなすと，T 個のソーシャルセンサが機能しているといえる．なお，発言に付与されているユーザ ID を参照すればユーザ数を求めることも可能ではあるが，リアルタイム性を重視したシステムへの実装には向いていない．そのため，発言数をそのままユーザ数と比例するものとみなしている．

ピーク中：インフルエンザが流行し始めると，ニュースなど，遠く離れた場所からでも対象地域のインフルエンザについての言及(遠距離言及発言)が増える(この数を IRT_{gap} とみなす)．この遠距離言及発言の量に応じて，話題としての価値が低減し，インフルエンザ関連発言を行わないユーザが増えるとみなす．このようなユーザはソーシャルセンサとして機能しないため，本稿では，**劣化ソーシャルセンサ**と呼ぶ．劣化ソーシャルセンサの数は，スケールパラメータ W を用いて，遠距離言及発言数を log で鈍らせた $W \cdot \log(IRT_{gap}+1)$ とする．なお，発言数(話題の度合い)は，対象地域によっては極端に大きな発言数となることもある非常に偏った分布(べき分布に近い)になる可能性があるため，べき分布を扱う際に一般的な log による対数をとっている．

このように考えると，対象地域 a の特定の日 t における患者数($I_1(a,t)$)(なお，単位は数ではなく対数とする)は以下のようにモデル化できる．

$$I_1(a,t) = \bar{I_1} \cdot \frac{M(a,t)}{N(a) - \bar{I} \cdot \log(G(a,t)+1)} \tag{4.1}$$

ここで，$M(a,t)$ は対象地域 a における任意の日 t におけるインフルエンザ関連発言数，

$G(a,t)$ は対象地域外からの対象地域名(県名)を含むインフルエンザ関連発言数，$N(a)$ は対象地域 a のソーシャルセンサ数(式(2.1)と同様に，対象地域における対象期間中の平均発言数)，$\log(G(a,t)+1)$ は話題の度合いを表す．なお，$G(a,t)$ は任意の日 t における対象地域 a に関するニュースや RT の発言量であり，インフルエンザ患者が発信するツイート数よりも指数的に増加すると考えられる．そのため，対数を用いてこのような発言による影響を抑えている．また，$\bar{I}_1$ と $\bar{I}$ はそれぞれスケールパラメータである．実験では，評価データにより患者数との誤差が最小となるようにフィッティングを行い，$\bar{I}_1$ は 1，$\bar{I}$ は 20 とした．なお，今回はツイート数と現実の統計量との差異をなるべく小さくするようなアプローチに焦点を当てており，パラメータやパフォーマンスの最適化についての検討は，今後の課題である．

このモデルは，ピーク前とピーク後などピークのタイミングを必要とせず，遠距離言及発言の数により，ピーク前後 2 つの状態を再現できる．すなわち，遠距離言及発言の数により，徐々にソーシャルセンサの数が低下し，この結果，インフルエンザ関連発言数の価値が相対的に高まる．

4.2 結果

提案手法(PROPOSED)の精度を測るために，感染症情報センターの報告をゴールドスタンダードとして，相関係数を求めた．比較のために，先行研究(EMNLP2011)(Aramaki et al., 2011)との相関係数を用いた．なお，先行研究(EMNLP2011)における患者数(I_0)推定式は式(2.1)の通りである．

2012–2014 年の結果を表 1 に示す．この結果に示されるように，全シーズンを通して PROPOSED の相関係数が EMNLP2011 の相関係数を上回っていることが分かる．都道府県別での相関係数を図 7 に示す．ほぼすべての地域で提案モデルの方が EMNLP2011 よりも相関係数が高く，精度が向上している．

提案モデルは，単純なものであるが，「話題としての価値が低くなると，発言量が減る」という人間の性質を取り込むだけで，2 章で述べたピーク後の発言量が実際よりも小さくなるという誤差を補正できる．このことは，今後，Twitter などのソーシャルメディアデータをより深く活用する際の重要な知見であると考える．

表 1．年度別の感染症情報センターの報告との相関係数の比較．

Method	2012	2013	2014	TOTAL
PROPOSED（劣化ソーシャルセンサモデル）	0.79	0.73	0.73	0.74
EMNLP2011	0.74	0.68	0.67	0.69

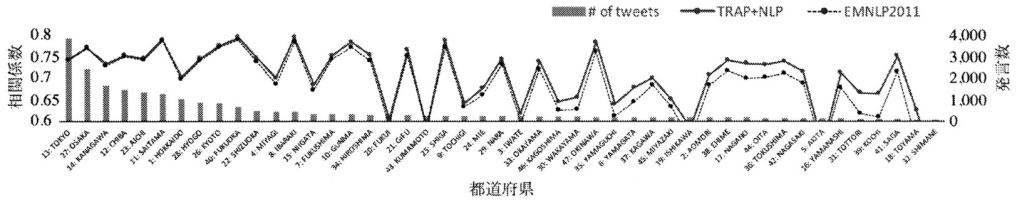

図 7．地域別の感染症情報センターの報告との相関係数の比較．X 軸は都道府県(並び順は各地域の Twitter 発言数(棒グラフ)に基づく)．Y 軸は，感染症情報センターの報告との相関係数．実線は提案手法(PROPOSED)，破線は先行研究(EMNLP2011)を示す．

5. 関連研究

Twitter をはじめとしたソーシャルメディアの普及により，新しくいくつかの研究が始まった，その代表的な例として 2 つの研究を示す．一つは，ソーシャルメディアの非文法的でノイジーな文章に対して処理の頑健性を高める研究である．この結果，Web 上にあるような非文法的な文章に対して，固有表現の抽出や単語の正規化を行う研究が行われてきた（Chrupała, 2014; Han and Baldwin, 2011; Plank et al., 2014）．

もう一つは，ソーシャルメディアから現実の世の中の知識を抽出する研究である．先の研究が自然言語処理に対する新たな課題であるとすると，こちらは新たな応用であるといえる．この結果，ソーシャルメディアからの意見抽出（O'Connor et al., 2010），イベント抽出（Li et al., 2014a; Marchetti-Bowick and Chambers, 2012; Sakaki et al., 2010; Shen et al., 2013; Thelwall et al., 2011），ユーザ行動分析（Bergsma et al., 2013; Han et al., 2013; Li et al., 2014b; Zhou et al., 2014），災害対応（Varga et al., 2013），世界知識抽出（Williams and Katz, 2012）など，様々なアプリケーションが提案されてきた．これらの応用例の中でも，疾患情報（特に，即時的な把握が必要とされる感染症）の流行検出に関しては，主要な Twitter 利用法の一つとして多くの研究がある（Aramaki et al., 2011; Paul and Dredze, 2011; 谷田 他, 2011）．Paul and Dredze (2011) は，病名ラベルが付与された発言なしで，より幅広い病気に関する発言を抽出する手法を提案している．そのために，事前知識として病気について書かれた記事を利用して Ailment Topic Aspect モデル拡張を行っている．これに対し，本研究ではインフルエンザに対象を絞り，発言のみを用いてインフルエンザ罹患者の判定を行い，患者数を推定している点が異なる．谷田他 (2011) は，風邪の流行度合いを推測するために，発言を用いて風邪の流行と関連した単語の出現頻度を回帰分析している．そのための変数選択において，選択する単語同士の相関をもとに，風邪の流行の特徴を捉えた推測を可能としている．一方，本研究はインフルエンザを対象にして，インフルエンザ罹患者による発言であるか否かを判定しているため，単純にあらかじめ決めた単語（今回は「インフル」）を用いるだけでも，都道府県単位の患者数を高い精度で推定できることを示している．

このような研究が盛んに行われているのは，感染症は未だ百万人を越える患者を出しており，恒常的な対策が必要であること（国立感染症研究所, 2006），および，新型インフルエンザといった危機事象についても，危惧されているという現状があるからである（Ferguson et al., 2005）．このため，感染症流行の把握は**感染症サーベイランス**と呼ばれ，各国で膨大なコストをかけて調査・集計が行われている．本邦でも，2016 年度から，国立研究開発法人日本医療研究開発機構（AMED）にて研究班が立ち上がり，Twitter のようなユーザ投稿発言データの利活用が進みつつある（国立研究開発法人日本医療研究開発機構（AMED）, 2015）．これにともない，本稿にて Twitter を中心に述べたソーシャル・メディアと現実の差異に関する議論が今後より進むものと思われる．

6. おわりに

本研究では時間的，および空間的ずれについて示した．いずれにおいても，人間の記述する欲求の偏りにより，不正確さを生んでいると考えられる．本研究で強調したいのは，ソーシャルメディアからの情報抽出に関する研究は，人間をセンサとみなしている（Sakaki et al., 2010）ものの，それはムラの多いセンサであることである．これを使いこなすためには，センサとしての人間の性質を十分に理解し，解析する必要がある．時間的には人々のピーク前の関心の加熱により，空間的には個々の地名の特徴（広さ，施設の個数）により，複雑な現象が生じ，ずれを生んでいる．これらを説明するためには，今後，Web において発言する人間の心理を真の対

象として解析することが必要となる可能性がある．同時に，人間の心理については，これまで心理学や社会学の分野で多くの知見が集積されているが，今後ソーシャルメディアの解析において，これら人間の心理を解析した知見との融合が必要になると考えられる．

参 考 文 献

Antoine, Émilien, Jatowt, Adam, Wakamiya, Shoko, Kawai, Yukiko and Akiyama, Toyokazu (2015). Portraying collective spatial attention in Twitter, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*), 39–48.

Aramaki, Eiji and Wakamiya, Shoko (2016). NAIST-ARS Guideline Ver. 1, https://dx.doi.org/10.6084/m9.figshare.3123160.v1 (in Japanese).

Aramaki, Eiji, Maskawa, Sachiko and Morita, Mizuki (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1568–1576.

Awamura, Takashi, Kawahara, Daisuke, Aramaki, Eiji, Shibata, Tomohide and Kurohashi, Sadao (2015). Location name disambiguation exploiting spatial proximity and temporal consistency, *Proceedings of the International Workshop on Natural Language Processing for Social Media* (*SocialNLP*), 1–9.

Bergsma, Shane, Dredze, Mark, Van Durme, Benjamin, Wilson, Theresa and Yarowsky, David (2013). Broadly improving user classification via communication-based name and location clustering on Twitter, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 1010–1019.

Bollen, Johan, Mao, Huina and Zeng, Xiaojun (2011). Twitter mood predicts the stock market, *Journal of Computational Science*, **2**, 1–8.

Chrupała, Grzegorz (2014). Normalizing tweets with edit scripts and recurrent neural embeddings, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 680–686.

Ferguson, Neil M., Cummings, Derek A. T., Cauchemez, Simon, Fraser, Christophe, Riley, Steven, Meeyai, Aronrag, Iamsirithaworn, Sopon and Burke, Donald S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature*, **437**(7056), 209–214, http://www.ncbi.nlm.nih.gov/pubmed/16079797.

Han, Bo and Baldwin, Timothy (2011). Lexical normalisation of short text messages: Makn Sens a #twitter, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, 368–378.

Han, Bo, Cook, Paul and Baldwin, Timothy (2013). A stacking-based approach to Twitter user geolocation prediction, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: *System Demonstrations*, 7–12.

国立感染症研究所 (2006).『インフルエンザ・パンデミックに関する Q&A（2006.12 改訂版）』, 国立感染症研究所 感染症情報センター, 東京.

国立研究開発法人日本医療研究開発機構（AMED）(2015). 平成 28 年度「新興・再興感染症に対する革新的医薬品等開発推進研究事業」に係る公募について, http://www.amed.go.jp/koubo/010620151113-01.html.

Li, Jiwei, Ritter, Alan, Cardie, Claire and Hovy, Eduard (2014a). Major life event extraction from Twitter based on congratulations/condolences speech acts, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 1997–2007.

Li, Jiwei, Ritter, Alan and Hovy, Eduard (2014b). Weakly supervised user profile extraction from Twitter, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 165–174.

Marchetti-Bowick, Micol and Chambers, Nathanael (2012). Learning for microblogs with distant su-

pervision: Political forecasting with Twitter, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 603–612.

奈良先端科学技術大学院大学ソーシャル・コンピューティング研究室 (2016). INFLU-KUN: Twitter-based Influenza Surveillance System, http://mednlp.jp/influ/.

O'Connor, Brendan, Balasubramanyan, Ramnath, Routledge, Bryan R. and Smith, Noah A. (2010). From Tweets to polls: Linking text sentiment to public opinion time series, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (*ICWSM*), 122–129.

odomon.net (2013). Twitter ユーザー数[2013 年第一位 東京都], http://todo-ran.com/t/kiji/13528.

Pang, Bo, Lee, Lillian and Vaithyanathan, Shivakumar (2002). Thumbs up?: Sentiment classification using machine learning techniques, *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, 79–86.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analysing Twitter for public health, *Processing of the Fifth International AAAI Conference on Weblogs and Social Media* (*ICWSM*).

Plank, Barbara, Hovy, Dirk, McDonald, Ryan and Søgaard, Anders (2014). Adapting taggers to Twitter with not-so-distant supervision, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*: *Technical Papers*, 1783–1792.

Sakaki, Takeshi, Okazaki, Makoto and Matsuo, Yutaka (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors, *Proceedings of the 19th international conference on World Wide Web* (*WWW*), 851–860.

Shen, Chao, Liu, Fei, Weng, Fuliang and Li, Tao (2013). A participant-based approach for event summarization using Twitter streams, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 1152–1162.

谷田和章, 荒牧英治, 佐藤一誠, 吉田稔, 中川裕志 (2011). Twitter による風邪流行の推測, TETDM&情報編纂研究会(第 6 回), 42–47.

Thelwall, Mike, Buckley, Kevan and Paltoglou, Georgios (2011). Sentiment in Twitter events, *Journal of the American Society for Information Science and Technology*, **62**(2), 406–418.

Tumasjan, Andranik, Sprenger, Timm O., Sandner, Philipp G. and Welpe, Isabell M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (*ICWSM*), 178–185.

Varga, István, Sano, Motoki, Torisawa, Kentaro, Hashimoto, Chikara, Ohtake, Kiyonori, Kawai, Takao, Oh, Jong-Hoon and De Saeger, Stijn (2013). Aid is out there: Looking for help from tweets during a large scale disaster, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 1619–1629.

Williams, Jennifer and Katz, Graham (2012). Extracting and modeling durations for habits and events from Twitter, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 223–227.

Zhou, Deyu, Chen, Liangyu and He, Yulan (2014). A simple Bayesian modelling approach to event extraction from Twitter, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (*Volume 2*: *Short Papers*), 700–705.

246 Proceedings of the Institute of Statistical Mathematics Vol. 64, No. 2, 233–246 (2016)

Difference between Number of Tweets and Real World Statistics

Eiji Aramaki and Shoko Wakamiya

Nara Institute of Science and Technology (NAIST)

The prevalence of social media services has brought a new approach for surveying people and social conditions. So far, various systems, such as an influenza surveillance system, an earthquake detection system and so on, have been proposed. However, information shared on social media doesn't always correspond to the real one. For example, social media services often suffer from rumors, causing lower reliability than existing media. In addition, several studies have been pointed out a limitation of both temporal and spatial accuracy in social media services. In this paper we examine the differences in terms of temporal and spatial perspectives based on Twitter data collected using our influenza surveillance system. Furthermore, we discuss a bias behind the differences.

Key words: Social media, Twitter, natural language processing, social computing, influenza.