Vol.2016-NL-227 No.17
2016/7/30

複数時点の単語出現頻度を扱う時系列データモデリング

磯 颯[1,a)]　若宮 翔子[1]　荒牧 英治[1]

概要：ソーシャルメディアの普及に伴い，様々な情報がインターネット上で共有されている．この結果，様々な社会現象および自然現象をインターネット上の情報から把握できるようになっている．特に，感染症に関するサーベイランス（現状把握技術）は，かつてない即時性から大きく注目されている．本研究では，ソーシャルメディア上で話題として取り上げられることが最も多い感染症の一つであるインフルエンザを題材に，従来のような現状把握だけでなく，流行予測を行うことを目指す．まず，実際に感染症が流行する前に，感染症の予防に関する情報が共有されていることに注目し，インフルエンザの流行を早期に示すような語を自動的に検出する．次に，実際の流行と任意の単語の相互相関係数を計算し，適切な時間ギャップの分だけタイムシフトした単語頻度を用いてモデルを構築し，患者数の予測を行う．この相互相関係数によるタイムシフトは，現状予測モデルの自然な拡張であるとととともに，インフルエンザのみならず，あらゆる感染症の予測に適応可能な手法である．2012 年 8 月から 2016 年 1 月までの，インフルエンザに関連する 770 万発言を用いた実験の結果，現状の患者数を相関係数平均 0.93 で推定し，1 週間先の患者数を相関係数平均 0.91，3 週間先の患者数を相関係数平均 0.76 で予測することができた．この結果は，現状推定については，本邦における最高精度である．予測については，初めての試みであり，今後の適応範囲の拡大，および，実用化が望まれる．

キーワード：Twitter，感染症サーベイランス，患者数予測，時系列モデリング，高次元データ解析

1. はじめに

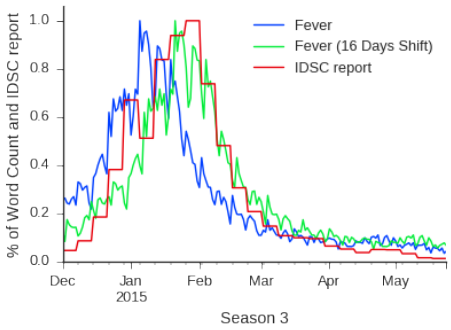
Twitter に代表されるソーシャルメディアの普及により，かつてない膨大な量の情報が発信・共有されている．これに伴い，様々な社会現象および自然現象をソーシャルメディアから把握する試みに注目が集まっている．こうした一連の研究は，ソーシャルセンサーと呼ばれ，地震の検出 [1]，感染症の広がり [2] や株価の変動推定 [3] などを中心に有効性が確かめられてきた．中でも，本研究では，代表的な感染症であるインフルエンザを扱う．ソーシャルメディアでインフルエンザを扱う研究は多く，レビュー論文によると，30 報以上もの研究があり，情報科学のみならず，医学領域での研究も多い [4]．このように，多数の研究が存在するソーシャルメディアベースのインフルエンザ・サーベイランスであるが，その多くに共通している仮定がある．それは，「Twitter 上でのインフルエンザの広がりが，現実世界のインフルエンザの広がりを直ちに反映している」ことである．この仮定の結果，多くの研究が，インフルエンザに関連する発言を収集し，一定時間単位（例えば，1 日）ごとに数え上げることで，実際の患者数を推定している．しかし，実際の発言には以下のようなものが多い：

- “インフルかかったので，今週は学校行かない🛌”
- “熱めっちゃ上がってる🤒 インフルかなぁ…”

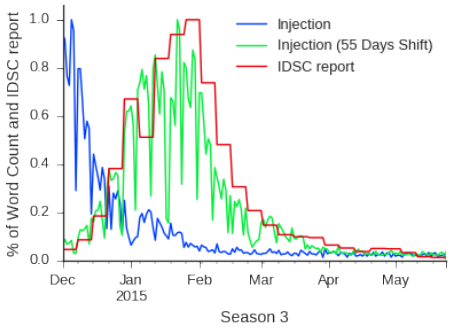
前者はインフルエンザに罹患した患者の発言であると推測される（インフルエンザの事実性がある発言）．一方，後者は，熱の事実はあるものの，インフルエンザについては疑いの域を出ない（インフルエンザの事実性がない発言）．このような両者の区別の問題は，事実性の問題として，前者のみを残し，後者を無視する分類アプローチが研究されてきた [5], [6].

しかし，後者は単なる非事実として無視してよいものであろうか？ 実際には，インフルエンザに罹患する直前に，人々は「熱」や「頭痛」といった語を含む発言をする傾向がある．例えば，図 1a に，「熱」という語の頻度の推移と実際のインフルエンザ患者数推移を示す．両者の増減パターンは類似しているが，それぞれのピークには時間ギャップがあり，「熱」の頻度のピークは実際のインフルエンザ患者数のピークよりも約 16 日早いことがわかる．このような時間ギャップがある発言を，単に非事実として無視する従来のアプローチよりも，将来的な事実の可能性として，時間ギャップを考慮してカウントするアプローチが有効である

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[a)] iso.hayate.id3@is.naist.jp



(a) 「熱」



(b) 「予防」

図 1: 各語の頻度と実際の患者数の推移.

と考えられる．具体的には，語の頻度を時間ギャップ分だけタイムシフトすれば，実際の患者数推移と適切にフィットさせることができる．図 1a の例では，「熱」の頻度を 16 日後ろにタイムシフトすることで，実際の患者数に近似させることが可能となる．

同様な患者数推移との時間ギャップは，「熱」や「頭痛」以外の単語にも存在する．それだけでなく，以下のように，感染症予防に関する単語全般について，時間ギャップを考慮したタイムシフトにより，患者数推移とフィットさせることができる．

- “インフルの予防接種行ってきた👍”
- “注射💉も嫌だけど，インフルになるのも嫌だ😩”

この例では「予防」と「注射」が該当する．図 1b に「予防」という語の頻度推移と実際の患者数推移を示す．この例では「予防」の頻度に 55 日もの長期間のタイムシフトを行うことにより，実際の患者数と最も高い相関を得ることができる．

本研究では，こうした語の頻度推移と患者数推移との相関が最も高くなる時間ギャップを相互相関係数により推定する．この時間ギャップに相当する時間をタイムシフトすると，単語頻度と患者数モデルとの相関はより高くなるはずである．そこで，全ての単語頻度をタイムシフトした上でモデル化することにより，より高い精度で患者数を推定する．この提案モデルにより，患者数との相関係数が平均 0.93 となり，タイムシフトを行わないモデルに比べ，大きく精度を向上することができた．

さらに，提案モデルは，タイムシフト先を未来にすることで，予測モデルへの拡張を容易にかつ自然に実現できる．提案モデルは，現在を日時 t とし，最大のタイムシフト幅を $\tau_{\max}$ とすると，過去 (day $t-\tau_{\max}$) から現在（day t）までの単語頻度をに基づき，現在（day t）の患者数を予測する．これを Δf 日未来にシフトし，過去（day $t-\tau_{\max}+\Delta f$）から現在（day t）までの単語頻度で，未来 (day $t+\Delta f$) の患者数を予測するように変形できる．実際に収集した発言を用いて行った実験では，7 日先の患者数予測精度の平均相関係数が 0.91，21 日先の患者数予測精度の平均相関係数が 0.76 となり，一定の実用化の目処となる相関係数 0.80 に近い精度で 3 週間も先の患者数を予測可能であることを示した

2. データセット

インフルエンザに関する発言のコーパス

Twitter API を用いて 2012 年 8 月から 2016 年 1 月までの，インフルエンザに関連する 770 万発言を収集した [*1]．インフルエンザの集計にあたっては，ノイズとなりうる発言として，リツイート（*RT*）を含む発言（本人以外の発言）やテキストリンク “*http*” を含むような発言（ニュースなどへの引用／言及が多い）を除去する．日本語形態素解析器としては JUMAN[*2] を用いて，解析結果の代表表記を用い形態素単位（以降，単に語と呼ぶ）の，頻度を集計する．低頻度語（頻度が 5 未満）を除いた結果，27,588 語の日毎の頻度推移行列が得られた．

IDSC による報告 （Gold Standard）

日本では，国立感染症研究所 (Infectious Disease Surveillance Center: IDSC) が，インフルエンザ流行期 (主に 11 月から 5 月の間) に，協力医療施設からの報告を取りまとめたインフルエンザ患者数を週 1 回のペースで報告している．本研究では，これを評価に用いることとし，以下のように 3 つの実験期間を設定した．2012/12/01 - 2013/05/31 (Season 1), 2013/12/01 - 2014/05/31 (Season 2), 2014/12/01 - 2015/05/24 (Season 3).

なお，実験に用いるタイムシフト幅を確保するため，実験期間の直前にバッファ期間（タイムシフト準備期間）を用意している．

3. 手法

本研究では，インフルエンザ患者数を，以下の式のように，複数の語頻度の線形結合でモデル化する．

$$\hat{y}^{(t)} = x_1^{(t-\hat{\tau}_1)}\hat{\beta}_1 + x_2^{(t-\hat{\tau}_2)}\hat{\beta}_2 + \cdots + x_{|V|}^{(t-\hat{\tau}_{|V|})}\hat{\beta}_{|V|},$$

[*1] Twitter API の仕様変更などの事情により，2013 年の 7 月から 10 月と 2014 年の 7 月から 10 月にかけて，発言を取得することができなかった．

[*2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

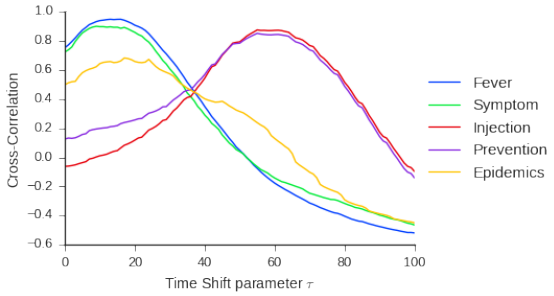


図 2: 5 つの語の頻度と IDSC 報告との相互相関係数.

ここで，t はある日時，$\hat{y}^{(t)}$ は日時 t における推定患者数，$x_v^{(t)}$ は語 v の日時 t における発言頻度，$\hat{\beta}_v$ は，語 v に対し推定された線形モデルの重みパラメータを表す．また，$\hat{\tau}_v$ は語 v に対し推定されたタイムシフト幅を表すパラメータである．

本章では，まず，語毎に最適なタイムシフト幅を推定する手法を説明し (3.1 節)，次に，タイムシフトを施した語頻度を用いた現状把握モデルを導入する (3.2 節)．最後に，この現状把握モデルを予測モデルへ拡張する (3.3 節)．

3.1 タイムシフト幅の推定

実際の患者数に対し，各語の頻度をどれだけタイムシフトしたときに最も強く相関するのかを知る必要がある．本研究では，その指標として，相互相関係数という評価指標を導入する．相互相関係数は，以下の式で定義される．

$$r_{\mathbf{x}_v,\mathbf{y}}(\tau) = \frac{\sum_{t=1}^{T}(x_v^{(t-\tau)} - \bar{x}_v^{(t-\tau)})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_v^{(t-\tau)} - \bar{x}_v^{(t-\tau)})^2 \sum_{t=1}^{T}(y_t - \bar{y})^2}},$$

ここで，τ はタイムシフト幅を表す．相互相関係数により，τ 日分タイムシフトしたときの語の頻度と実際の患者数との相関を見ることができる．

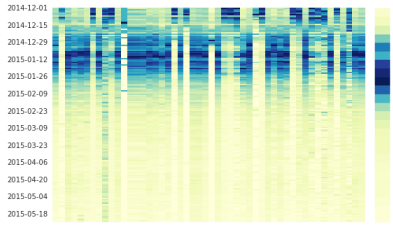
図 2 は，ある語に対して，τ を 0 から 100 の間で，それぞれ相互相関係数を計算したときの値の推移を表している．例えば，インフルエンザ患者が頻繁に発言する傾向にある「熱」や「症状」といった語の相互相関係数は，16 日付近で最も大きくなっていることが分かる．一方で，「予防」や「注射」といった語の場合，55 日程度タイムシフトすると実際の患者数と強く相関することが分かる．各語に対するタイムシフト幅として，最も相互相関係数が大きくなる日数を用いる．得られた語群を用いて線形モデルの推定を行うために，タイムシフトを施した語の頻度を行列の形で得る必要がある．行列の構築には Algorithm 1 を用いる．

具体的には，各語に対して，インフルエンザ患者数と相互相関係数を計算し，指定したタイムシフト範囲の中で [*3]，最も大きな相互相関係数が得られるタイムシフト幅を探索する．次に，全ての語に対して，タイムシフトを施し，新たな語頻度ベクトルを得る．この結果，全ての単語頻度がタイムシフト幅の範囲で，インフルエンザ患者数と最大の相関となる行列が得られる．

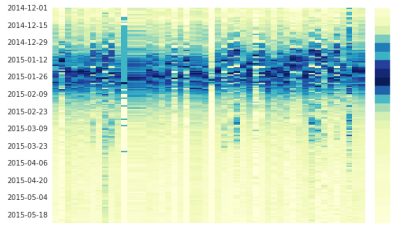
[*3] データセットの欠損期間を考慮し，本研究では，最大のタイムシフト幅 $\tau_{\max}$ を 60 とした．

Algorithm 1: Time Shift Word Matrix

```
Set maximum shift parameter τ_max
for v ← 1 to |V| do
    for τ ← 0 to τ_max do
        Calculate Cross Correlation r_{x_v,y}(τ)
    end
    τ̂_v = argmax_{τ∈{0,...,τ_max}} r_{x_v,y}(τ)
    Shift word vector to maximize Cross Correlation
    x̂_v ← [x_v^{(1−τ̂_v)}, x_v^{(2−τ̂_v)}, ..., x_v^{(T−τ̂_v)}]
end
return Shifted Word Matrix X = [x̂_1, ..., x̂_{|V|}]
```



(a) タイムシフトなしの語の行列.



(b) タイムシフトありの語の行列.

図 3: 頻度推移行列のタイムシフト.

図 3 にこのアルゴリズムの概念図を示す．各図の左側の行列が全語彙中から選択した 50 語の頻度推移，右側のベクトルが IDSC 報告を表しており，それぞれ最大値を取るところで色の濃さが最大となる．図 3a はタイムシフト前の語の行列である．この図から分かるように，実際の患者数の傾向よりも多くの語が先にピークを迎えており，さらに，見かけ上は複数の語が全く関連していないように見える．しかし， 図 3b に示すように，タイムシフトを施した行列は，実際の患者数と非常に似た傾向を示していることが分かる．

3.2 パラメータ推定

タイムシフトされた単語頻度行例から，線形モデルを構築するために，語毎のパラメータ β を推定する必要がある．しかし，我々のデータは，語のサイズ $|V|$ が，実験日数 T よりも非常に大きくなっているため，最小二乗法による推定ではパラメータの解が一意に得られず，推定結果

も学習データに過剰に適合した値を返してしまう可能性がある．したがって，以下のようにパラメータに制約を加えた推定を行う：

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mathbf{P}(\boldsymbol{\beta}, \boldsymbol{\lambda}),$$

ここで，$\boldsymbol{P}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ は制約項を表す．制約項については，$\mathbf{P}_{lasso}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \lambda \|\boldsymbol{\beta}\|_1$ とした Lasso をはじめとし，多くの研究がある [7]．本研究では，先行研究 [8] にて用いられた Lasso とともに，特徴量の数がサンプル数よりも多く，特徴量同士が強く相関しているような場合に対して頑健な Elastic Net を用いた [9]．Elastic Net は，l_1 正則化と l_2 正則化を組み合わせた次のような制約項 $\mathbf{P}_{enet} = \lambda(\alpha \|\boldsymbol{\beta}\|_1 + (1-\alpha) \|\boldsymbol{\beta}\|_2^2)$ を持つ．ここで，α を l_1 比と呼ぶ．$\alpha = 1$ ならば，Elastic Net は Lasso と等価であり，$\alpha = 0$ のとき，Ridge (l_2 正則化) と等価である．Elastic Net も Lasso のように，特徴選択と縮小推定を同時に行い，l_2 正則化の利点である，非常に似通った特徴量を同時に選択するような特徴を持つ．Elastic Net は，Lasso と Ridge のいずれも内包しており，両者と同等かそれ以上によいモデルを推定できることが期待できる．

3.3 予測モデル

提案モデルは，自然に予測モデルへの拡張が可能である．例えば，Δf 日先の患者数を予測することを考える．この場合，各語のタイムシフト幅を探索する際，少なくとも Δf 日タイムシフトするという制約を課すことで，観測済みの語の頻度から Δf 日未来の予測を行うことができる．予測のための行列構築には，Algorithm 2 を用いる．

Algorithm 2: Time Shift Word Matrix for Prediction.

```
Set maximum shift parameter τ_min, τ_max
for v ← 1 to |V| do
    for τ ← τ_min to τ_max do
        Calculate Cross Correlation r_{x_v,y}(τ)
    end
    τ̂_v = argmax_{τ∈{τ_min,...,τ_max}} r_{x_v,y}(τ)
    Shift word vector to maximize Cross Correlation
    x̂_v ← [x_v^{(1−τ̂_v)}, x_v^{(2−τ̂_v)}, ..., x_v^{(T−τ̂_v)}]
end
return Shifted Word Matrix X = [x̂_1, ..., x̂_{|V|}]
```

4. 実験 1：モデリング

まず，先行研究と同様に，実際の患者数を Twitter から予測するモデルを推定する．

4.1 比較手法

我々は，以下の 4 つの線形回帰モデルを用いて，それぞれの精度比較を行う．

- `Lasso`: l_1-正則化法 [7], [8],
- `Lasso+`: 時間ギャップを考慮したデータを用いた Lasso
- `ENet`: l_1-, l_2-正則化を組み合わせた手法 [9],
- `ENet+`: 時間ギャップを考慮したデータを用いた Elastic Net

全てのハイパーパラメータは，5-fold cross validation により決定した．

4.2 データと評価指標

使用するデータに関する詳細は 2 章を参照されたい．これを 3 つの期間に分割し，各々を学習データとして，学習に用いなかったデータを評価用のデータとして，それぞれ検証を行った．

なお，評価には，モデルの推定値と感染症情報センター報告（以降，IDSC 報告）[*4] の患者数との相関係数を用いた．

4.3 結果

各モデルの推定値と IDSC 報告との比較を図 4，実際の精度を表 1 に示す．図の縦軸は患者数，横軸はテストの日時を表している．ベースラインとして用いた `Lasso` や `Enet` は，先行研究と比較してもあまり高い精度は得られなかった．これは，今回のデータセットが，比較的難しいデータセットであることを示唆している．

一方で，時間ギャップを考慮したタイムシフトモデルは，ベースラインのモデルに比べ約 0.1 ポイント相関が大きくなり，モデリングの精度がより高くなることが分かった．なお，今回は，“*RT*” や “*http*” を含む発言を取り除くという単純な前処理を行っただけであるため，多くのノイズが残っている可能性がある．提案手法がこれらのノイズ除去を自然に行えているのか，または，ノイズ除去を行うとさらなに，精度がよくなるのか，今後，先行研究 [6] との比較・検証を行いたい．

さらに，Lasso と Elastic Net を用いて，それぞれのデータの推定をした結果，大きな違いは見られなかった．これは，Elastic Net を推定した際，l_1 比の値が 1 に近く，ほぼ Lasso と同様の推定を行っていたことに起因している．

過剰推定

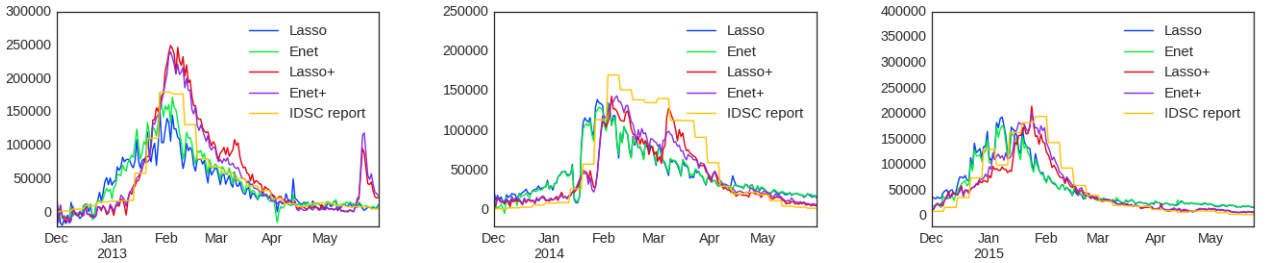
図 4a のように，実際の患者数に比べ，患者数を多く見積もっている場合が存在する．Season 1 では，“**「鳥＋豚インフルの合成ウイルス」作成される**”[*5] というニュースが話題となったことが影響し，過剰に患者数を多く推定している．前処理として，“*http*” を含む発言を予め除去していたが，それでも，多くのユーザが以下のようにニュースに関する話題を（ニュースへのリンクなしに）発言していた．

- “何？鳥インフルエンザと豚インフルエンザのワクチ

[*4] `http://www.nih.go.jp/niid/ja/from-idsc.html`
[*5] `http://wired.jp/2013/05/17/h5n1-h1n1-reassortment/`

表 1: 推定患者数と IDSC 報告との相関係数.

Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
Test	Season 1		Season 2		Season 3		
Lasso	0.854	0.916	0.768	0.894	0.770	0.753	0.820
Enet	0.900	**0.927**	0.809	**0.914**	0.792	0.805	0.823
Lasso+	**0.952**	0.907	0.951	0.888	0.955	**0.963**	**0.936**
Enet+	0.944	0.898	**0.960**	0.878	**0.967**	0.959	0.934



(a) Train: Season 2, Test: Season 1 (b) Train: Season 1, Test: Season 2 (c) Train: Season 1, Test: Season 3

(d) Train: Season 3, Test: Season 1 (e) Train: Season 3, Test: Season 2 (f) Train: Season 2, Test: Season 3

図 4: 4 つの手法それぞれの推定患者数と IDSC 報告との比較.

ンを作ろうとして新種のインフルエンザウィルスを作り出しちゃった？何やってるの？😡”

- “中国の研究者が，鳥インフルと豚インフルを組み合わせて新しいウイルスを作ったらしい😢”

ここで，Season 2 から構築されたモデルは，「鳥」という語を特徴量として一定の重みで選択していたため，大きな誤りが発生した．このような誤りを避けるために，突発的に発生する発言に対する対応が今後の課題である．

5. 実験 2: 予測

3.3 節で導入した予測モデルの評価を行った．

5.1 比較手法

タイムシフトを考慮しない Lasso や Enet は，将来の患者数を予測することはできない．したがって，Lasso+ と，Enet+ を予測モデルとして用いる．さらに，予測モデルのベースラインとして，以下のようなモデルを導入する．このベースラインは train と同じ流行を test で持つと仮定したものである．

BaseLine: $\hat{y}^{(t)}_{\text{test}} = y^{(t)}_{\text{train}}$

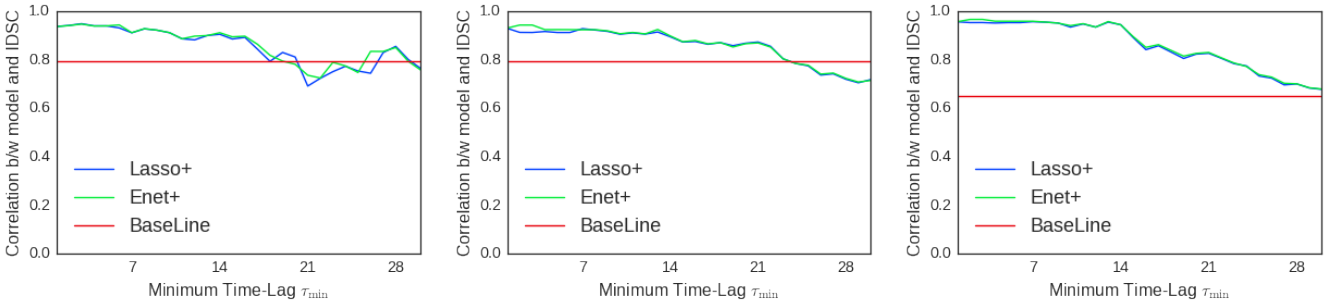
5.2 データセットと評価指標

予測モデルの評価には，実験 1 と同じデータと評価指標を用いた．また，タイムシフト幅 $\tau_{\min}$ を 1 日から 30 日までとした．

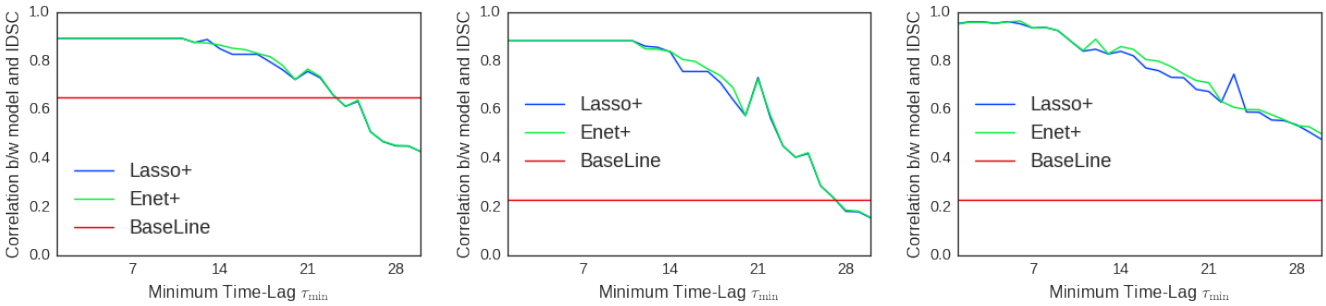
5.3 実験結果

実験結果を，図 5 に示す．各図の縦軸はモデルの推定患者数と IDSC 報告の相関，横軸は最小タイムシフト $\tau_{\min}$ を表している．両モデルとも，3 週間先の患者数に関しては，ベースラインよりも高い精度で推定することができた．また，1 週間先の患者数の予測は，実験 1 で求めた推定モデルとほぼ同じ相関が得られたことが分かる．これは，実験 1 で選択された多くの特徴量のタイムシフト幅が少なくとも 10 日ほどであったため，一週間先の患者数を推定するモデルでも同じような結果が得られたと考えられる．

図 6 では，予測モデルが実際に推定した患者数推移の典型的な例を挙げる．各図の縦軸は患者数，横軸はテストの日時を表している．図 6a から 図 6d は，Season 2 を学習データとして用い，Season 1 をテストデータとしたときの予測モデルの推定結果である．推定モデルは，一貫して実

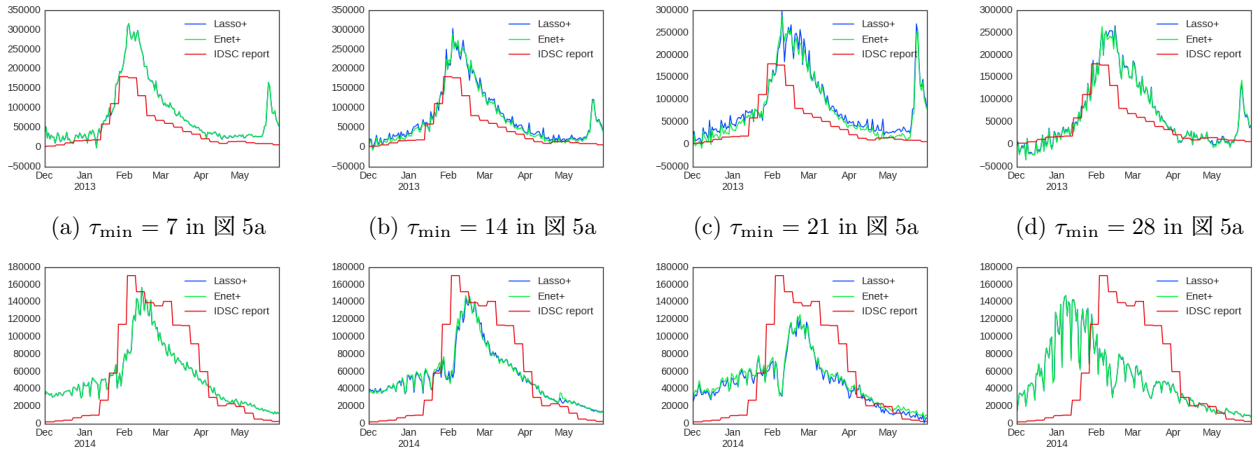


(a) Train: Season 2, Test: Season 1　(b) Train: Season 1, Test: Season 2　(c) Train: Season 1, Test: Season 3



(d) Train: Season 3, Test: Season 1　(e) Train: Season 3, Test: Season 2　(f) Train: Season 2, Test: Season 3

図 5: 各 $\tau_{\min}$ について `Lasso+` と `Enet+` を用いて予測した患者数と IDSC 報告との相関係数の遷移.



(a) $\tau_{\min} = 7$ in 図 5a　(b) $\tau_{\min} = 14$ in 図 5a　(c) $\tau_{\min} = 21$ in 図 5a　(d) $\tau_{\min} = 28$ in 図 5a

(e) $\tau_{\min} = 7$ in 図 5f　(f) $\tau_{\min} = 14$ in 図 5f　(g) $\tau_{\min} = 21$ in 図 5f　(h) $\tau_{\min} = 28$ in 図 5f

図 6: 各 $\tau_{\min}$ について `Lasso+` と `Enet+` を用いて予測した推定患者数と IDSC 報告との比較.

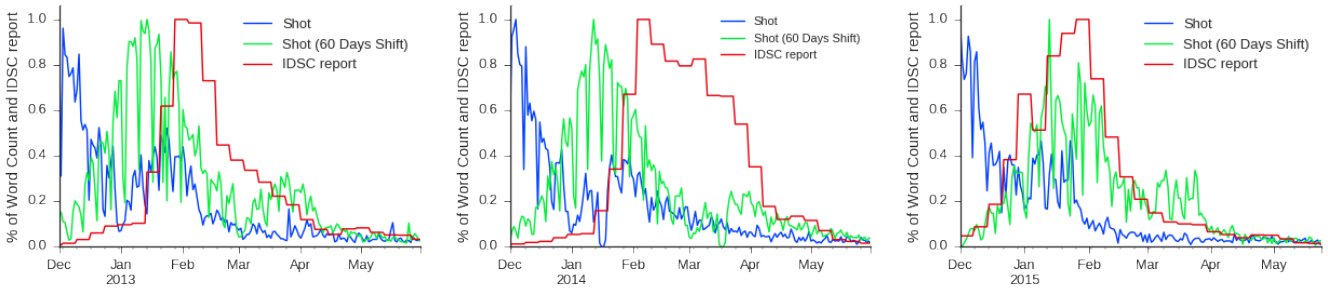
際の患者数と似た値を返していることが分かる．図 6c では，4.3 節で述べたような現象が顕著に現れている．一方，図 6d では，「鳥」という語にかかる重みが小さくなったため，図 5a で予測モデルの相関が 3 週間後よりも 4 週間後の方が高くなった．

また，図 6e から 図 6h では，Season 3 を学習データとして用い，Season 2 をテストデータとしたときの予測モデルの推定結果である．図 5e からもわかるように，他のケースと比べ $\tau_{\min}$ の増加に伴い，急激に予測精度が落ちていることが分かる．このモデルは，実際の流行が起こる前に，過剰に患者数を見積もっており，ピーク後は過少に患者数を見積もっている．特に，図 6h に示す $\tau_{\min} = 28$ のモデルで，その傾向が顕著に現れている．この現象に関して， 6 章でより詳しく言及する．

6. 考察

多くの場合において，提案した語の頻度に対するタイムシフトアプローチにより，IDSC 報告によりフィットする患者数の推定が可能となった．

一方で，各 Season において，同様なタイムシフトを持たない語が選択された場合，推定精度が落ちることも分かった．例えば，Season3 で学習した `Lasso+` モデルは，$\hat{\tau}_{\text{熱}} = 16$ としたときの語「熱」と，$\hat{\tau}_{\text{接種}} = 55$ としたときの「接種」，$\hat{\tau}_{\text{休む}} = 10$ としたときの「休む」といった語を



(a) Season 1　(b) Season 2　(c) Season 3

図 7: 各 Season における「打つ」の頻度推移.

それぞれ特徴として選択した．これらの語をタイムシフトした結果は，Season 3 の IDSC 報告と高い相関があった．

しかしながら，Season 1 や Season 2 においては相関が見られない語も存在した．その典型的な例の一つが，「接種」の $\hat{\tau}_{\text{接種}}$ である．実際に，「接種」は Season 3 の IDSC 報告と高い相関（0.850）が見られたが，Season 1 と Season 2 における相関はそれぞれ 0.235 と-0.04 となった．これは，Season 3 における $\hat{\tau}_{\text{接種}}$ が Season 1 や Season 2 と異なっていたためであると考えられる．この現象はタイムシフトなしのモデルから特徴量を選択する際にも起こりうる問題である．このように，一部の語については，安定した時間ギャップが存在していない場合がある．

特に，予測フェーズにおいては，時間ギャップの不安定さが精度に大きく影響する．図 5e において，予測精度は $\tau_{\min}$ の増加とともに，大きく減少した．$\tau_{\min} = 28$ とした際，Season 3 で学習したモデルは「打つ」という語に最も大きな重みを付与したにも関わらず，「打つ」という語の Season 1 と Season 2 における IDSC 報告との相関はそれぞれ 0.31 と 0.03 であった．実際に「打つ」のタイムシフト幅 $\hat{\tau}_{\text{打つ}}$ はすべての Season で，最大の 60 日となっているが，図 7 に示すように，Season 1 と Season 2 のいずれにおいても，より大きなタイムシフトを行ったほうが，実際の患者数によくフィットすることがわかる．このため，Season 3 で学習を行ったモデルを，Season 2 でテストする際，「打つ」にかかるべきタイムシフト幅が過少に推定されたため，推定モデルが実際のピークよりも早い段階で患者数を大きく見積もってしまった．このように，各 Season において，患者数との時間ギャップが異なる特徴量を選択してしまうと，予測精度は大きく減少する．

一方で，「熱」や「症状」などの語に関しては，各 Season において，同様な時間ギャップが存在していたため，近い将来の患者数であれば，高い精度で予測することができた．このように，今後，各語について，安定した時間ギャップを持つかどうかを判定することが必要となる．これは，今後，長期間にわたる観測を行うことで自然と可能であると思われる．

本研究の貢献をまとめる．まず，語頻度に対するタイムシフトは，インフルエンザ流行モデリングにおいて有効であることを確認した．さらに，既存研究では試されていなかった予測を行うことが可能となり，提案手法の拡張性の高さを示した．今後，時間ギャップが安定した質の高い語を選好したモデルを用いることで，より高い精度のモデリングや予測に取り組む予定である．

7. 結論

本研究では，ソーシャルメディア上で話題として取り上げられることが最も多い感染症の一つであるインフルエンザを題材に，現状把握だけでなく，流行予測を行うモデルを提案した．まず，実際に感染症が流行する前に，インフルエンザの流行を早期に示すような語を自動的に検出した．次に，実際の流行と任意の単語の相互相関係数を求め，適切な時間ギャップの分だけタイムシフトした単語頻度モデルを構築し，予測を行った．この相互相関係数によるタイムシフトは，現状予測モデルの自然な拡張であるとともに，インフルエンザのみならず，あらゆる感染症の予測に適応可能であると言える．インフルエンザに関連する 770 万発言を用いた実験の結果，3 シーズンにおいて，相関係数平均 0.93 で現状の患者数を推定し，相関係数平均 0.91 で 1 週間先の患者数を，相関係数平均 0.76 で 3 週間先の患者数を予測することができた．この結果は，現状推定については，本邦における最高精度である．予測については，初めての試みであり，今後の適応範囲の拡大，および，実用化が望まれる．

参考文献

[1] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, *Proceedings of the 19th International Conference on World Wide Web*, ACM, pp. 851–860 (2010).
[2] Chew, C. and Eysenbach, G.: Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak, *PLoS ONE*, Vol. 5, No. 11, pp. 1–13 (2010).
[3] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational*

ence, Vol. 2, No. 1, pp. 1–8 (2011).

[4] Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H. Y., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J. and Corley, C. D.: Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review, *PLoS ONE*, Vol. 10, No. 10, pp. 1–20 (2015).

[5] Kanouchi, S., Komachi, M., Okazaki, N., Aramaki, E. and Ishikawa, H.: Who caught a cold ? - Identifying the subject of a symptom., *Proceedings of The Association for Computer Linguistics*, pp. 1660–1670 (2015).

[6] Aramaki, E., Maskawa, S. and Morita, M.: Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter, *Proceedings of Empirical Methods in Natural Language Processing*, pp. 1568–1576 (2011).

[7] Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, Series B*, Vol. 58, pp. 267–288 (1994).

[8] Lampos, V. and Cristianini, N.: Tracking the flu pandemic by monitoring the Social Web, *Proceedings of the 2nd International Workshop on Cognitive Information Processing*, pp. 411–416 (2010).

[9] Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, Vol. 67, pp. 301–320 (2005).