ツイートされる病気・症状の可視化に向けた症状の事実性解析

安藤翼
香川大学大学院工学研究科
s19g453@stu.kagawa-u.ac.jp

安藤一秋
香川大学創造工学部
ando.kazuaki@kagawa-u.ac.jp

1 はじめに

近年，医療分野に自然言語処理を応用する研究[1,2,3]が注目されている．また，コロナ禍においては，新型コロナウイルス対策に AI 技術を応用する研究組織[4]が設立され，一般社会からも注目を浴びている．医療分野に応用する研究のうち，ソーシャルメディアを対象とした研究においては，Twitter からツイートを収集してインフルエンザの流行度合を推定し，その状況を可視化する研究[1]や，感染症のみを対象とし，病気の事実性を解析する研究[2]，インフルエンザと新型コロナウイルスに同時感染した場合における特徴・症状分析に関する研究[3]など，対象を特定の病気や症状に限定する研究が中心である．

本研究では，感染症であるか否かを問わず，いつ，どこで，どのような病気・症状がツイートされているのかを収集・分析し，地域別・時系列別に可視化するシステムの構築[5]を目的とする．本システムを実現するためには，既存研究[1,2]のような感染症名を含むツイートのみを対象とした事実性解析ではなく，病名を含まないツイートに対して病気・症状の事実性を解析する手法が必要となる．NTCIR-13 MedWeb Task[6]においては，8 つの病名又は症状を含むツイート文にアノテーションしたデータセットが公開され，これに基づく手法がいくつか提案されているが，対象とする病気・症状は少ない．

本稿では，Twitter に投稿された各ツイートに含まれる様々な症状に対して，二値分類により，事実性を解析する手法について検討する．

2 症状の事実性解析手法

2.1 データセットの構築

ツイートに含まれる症状に対して事実性解析する手法を検討するに先立ち，症状を含むツイートを Twitter 上から収集する必要がある．ツイートを収集するには，対象とする症状の辞書が必要となる．

本稿では，「goo ヘルスケア家庭の医学 [1)]」を基に対象とする症状を決定した．表 1 に，対象とする 14 種類の症状を示す．症状は，性別，年齢に関係しない一般的なものを採用した．

表 1　対象とする症状

頭痛	腹痛	発熱	下痢	充血
関節痛	吐血	めまい	嘔吐	動悸
しびれ	発疹	息切れ	寒気	

次に，実験に用いるデータセットを構築するためのツイート収集法について述べる．ツイートの収集条件は，ツイート内に 14 種の症状を必ず 1 語以上含むこととする．

以上の条件で収集した結果，合計 30 万件のツイートが得られた．その後，収集したツイートから無作為に 20,000 件を抽出し，ツイート内に含まれる症状が病気によるものなのか否かを精査し，人手で真偽ラベルを各ツイートに付与する．

以下，ラベル付与の 4 つの条件について示す．

① 対象者は当人及びその近辺者

ツイート内容が投稿者に対すること又は投稿者の近辺者に対する場合は「真」と判定する．ここで，収集したツイートには症状の対象者が記述されていない場合がある．そこで，「は」，「が」の 2 種の助詞が存在せず，先頭が症状名から始まらない文を含むツイートに対しては，1 人称代名詞として「私は」を自動で付与する．そして，上記以外の第 3 者である場合は「偽」と判定する．

② 時制は現在

ツイート内容から現在継続していると判断できる場合は 「真」，過去の話題と判断できる場合は「偽」と判定する．ただし，症状が過去から現在に至るまで継続していると判断できる場合は「真」と判定する．

[1)] 2021 年 1 月時点ではサービスを終了している．

③ 仮定表現

Twitter をセンサーとして利用するため，ツイート文内に「かもしれない」,「であろう」などの仮定発言がある場合も「真」と判定する.

④ 病気に関係する症状

ツイート内の症状が病気に関係するとき「真」と判定し, 気分や表現の拡張により症状の原因が読み取れない場合は,「偽」と判定する.

以上の 4 条件の下で，20,000 件のツイートに真偽ラベルを付与し，6,046 件の正例，13,954 件の負例からなるデータセットを構築した.

2.2 症状の事実性解析に利用する素性

我々の先行研究[7]では，Support Vector Machine（SVM），ロジスティック回帰（LR），多層パーセプトロン（MLP） の 3 種類の分類器と，Bag of Words，病名の有無，体の部位の有無，文長，単語分散表現の 5 つの素性を用いて，14 症状を対象とした事実性解析の評価を実施した．その結果，分散表現を素性とする SVM が最も分類性能が高く，0.743 という適合率を得ることを確認した.

本稿では，以下の 4 素性を用いたモデルの有効性について検討する.

① 単語分散表現 (W2V)

単語分散表現は，BoW 素性と異なり，単語の出現頻度だけでなく, 前後の単語の意味や関わりを考慮することができるため，単語分散表現を baseline 素性に利用する．分散表現は，株式会社ホットリンクが公開している「日本語大規模 SNS+Web コーパスによる単語分散表現モデル[8]」を用いる. このモデルは, SNS や Web 上の文書，日本語 Wikipedia を学習コーパスに利用している. 素性値には, 文内の形態素に対する分散表現の平均を利用する.

② つつじ[9]素性 (Ttj)

日本語機能表現辞書である「つつじ」を素性抽出に利用する.「つつじ」には,「に対して」や「かもしれない」などの機能表現とその意味が登録されている．文に含まれる機能をとらえることで，文の意味をより正確に学習できると考える. つつじ素性（Ttj）には，つつじ内の各機能表現に付与されている意味 ID を用いる．各文において，症状の右側最長 α 字中に含まれる最長の意味 ID を素性として利用する.

③ 病名素性 (Sick)

ツイート文内に病名が含まれている場合, 症状が病気に関係する可能性が高いと予測できるため, 文中に病名が含まれているか否かを素性として利用する．病名は「株式会社 メディカルノート[10]」に登録されている 2,102 種の病名を利用する.

④ 物体素性 (Object)

症状を含むツイート内において, 症状の対象が人であるとは限らない. 比喩や人以外の物体が対象となるツイート文も存在するため, 文内に人以外の物体が含まれているか否かを素性値として利用する．物体名には，真偽ラベルを付与した 20,000 件のツイートデータから人手で抽出した 29 種を利用する.

3 主体性判定手法

本稿では，症状の原因が病気であり，かつ，対象者がツイート投稿者又は，その近辺者である場合のみ，正例として扱う．そこで，2.2 項で検討した 4 素性を用いて，症状が原因であると判定されたツイートに対して，そのツイート対象者を分類し，ツイート投稿者又は，その近辺者であると判定されたツイートのみ可視化に利用する.

収集したツイートから，症状の事実性解析に使用したツイート以外を無作為に 20,000 件抽出し，症状の対象者に応じて，以下の 3 種のラベルを各ツイートに付与する.

① 当人・近辺者

② 遠方者

③ 不明

3.1 主体性判定に利用する素性

主体性判定では, 以下の 3 素性について検討する.

① 1 人称代名詞素性 (Pron)

ツイート内における症状の対象が投稿者当人である場合,「私」や「僕」などの 1 人称代名詞を記載する可能性が高い. そこで,「Wikipedia[11]」に登録された 30 種の公的表現及び私的表現に含まれる 1 人称代名詞がツイート内に含まれているか否かを素性に利用する.

② 都道府県素性 (Area)

ツイートの対象人物が遠方者である場合, どこに住む人物なのか，また，現在の居場所などの地名に関する情報が記述される傾向があると考える．そこで，ツイート内に都道府県名が存在する

場合，その都道府県名を素性に利用する．

③ 移動予測素性 (Move)

ツイートが, ある地点から別の地点へ移動する内容の場合, 対象人物が遠方者である可能性が高い. そこで, 移動を予想させる助詞がツイート中の名詞に続く場合, その助詞を素性として利用する．本稿で対象とする 7 種の助詞を以下に示す.

表 2　移動を予想させる助詞

に	へ	まで	から
で	ね	か	

以上の 3 種の素性に，2.2 項で提案した単語分散表現 (W2V)，つつじ素性 (Ttj) の計 5 種の素性を用いて，主体性を判定する．なお，主体性判定においても，単語の分散表現を baseline 素性とする．

4 評価実験

評価実験では，始めに，症状の事実解析を行い，病気が原因であると判定されたツイートに対して，そのツイートの対象人物を判定する．

4.1 実験方法

形態素解析には MeCab を利用する．また，先行研究[7]と同様，分類器には，Support Vector Machine (SVM)，ロジスティック回帰 (LR)，多層パーセプトロン (MLP)を用いる. 分類器の実装には, scikit-learn を使用し, 分類器ごとに, baseline 素性である単語の分散表現に他の検討素性を組み合わせて各々の分類性能を比較する．SVM には RBF カーネルを使用し, コストパラメータは 1,000 とした. MLP は最適化関数を SGD，バッチサイズを 128 とした．

4.2 評価方法

症状の事実性解析では，2.1 節で述べた 20,000 件のデータセットを用いる．適合率，再現率，F 1 値を評価尺度とし, 10 分割交差検証の平均値で分類性能を評価する．人手でラベルを付与した結果と分類結果を比較し，完全一致した場合を正解と判断する．なお，症状の罹患状況を可視化する際，病気が原因であると判定されたツイートのみ用いることを想定しているため，本実験では，適合率を最重視する．

主体性判定では，症状の事実性解析の結果，病気が原因であると判定されたツイートに対して，新たに 3 種の主体性ラベルを付与しテストデータとする．事実性解析に使用したツイート以外の 20,000 件に対し，3 種のラベルを付与すると共に， 2.1 項の方法で「私は」をツイートに自動付与したものを学習データに用いる．適合率，再現率，F1 値を評価尺度とし,「当人・近辺者」に分類された性能で評価する．本研究では，ツイートの対象者が当人又は近辺者である場合を正例として扱い,「遠方者」,「不明」に分類されたツイートは負例とする．人手でラベルを付与した結果と分類結果を比較し，完全一致した場合を正解と判断する．罹患状況の可視化には「当人・近辺者」へ分類されたツイートのみ用いることを想定しているため，適合率を最重視する．

4.3 実験結果

まず，症状の事実性解析において，3 種の分類器に beseline 素性及び 3 種の検討素性を組み合わせた結果を表 3 に示す.

表 3 より，最も適合率が高い組み合わせは，MLP に全ての検討素性を加えた場合で 0.788 を得た. 我々の先行研究[7]で提案した手法より，性能が 4.5 ポイント向上した．対して最も適合率が低くなった組み合わせは LR に baseline 素性である単語の分散表現のみを用いた場合で 0.732 となった. Ttj 素性のみに着目すると，全ての分類器において baseline 素性のみの場合と比較して, 分類性能が向上した. これは, ツイートの文末における意味的まとまりの表現を考慮することができたためであると考えられる. また, Object 素性のみを追加した場合も，全ての分類器において baseline 素性のみの場合と比較して，分類性能が向上した．ツイート文内に含まれる症状の対象は，人以外の「スマホ」や「パソコン」などの物体となる場合がある．そこで，ツイート内の物体に関する素性を追加することで，対象物の分類に貢献したと考えられる. 最後に, Sick 素性を追加した場合, 分類性能に大きな変化はなかった．これはツイート内に含まれる病名の件数が全体の 22%と低いことが原因であると予想できる．

次に，主体性判定において，3 種の分類器に対して，beseline 素性及び 3 種類の検討素性を組み合わせた結果を表 4 に示す．主体性判定では「当人又は近辺者」に分類されたツイートを最終結果とするため,「当人又は近辺者」に分類された結果のみを示す．また，表 4 の結果は，表 3 の症状の事実性解析の結果において，最も分類性能が高い MLP に全ての検討素性を加えたモデルが正例と判定した結果に

対し，主体性判定した結果である．

表 3　症状の事実性解析の結果

素性	適合率	再現率	F1
SVM (W2V)	0.749	0.749	0.749
SVM (W2V+ Ttj)	0.760	0.760	0.760
SVM (W2V+Sick)	0.748	0.747	0.747
SVM (W2V+Object)	0.754	0.754	0.754
SVM (All)	0.767	0.767	0.767
LR (W2V)	**0.732**	**0.730**	**0.730**
LR (W2V+Ttj)	0.766	0.767	0.767
LR (W2V+Sick)	0.733	0.733	0.733
LR (W2V+Object)	0.735	0.735	0.735
LR (All)	0.769	0.769	0.769
MLP (W2V)	0.746	0.746	0.746
MLP (W2V+Ttj)	0.773	0.774	0.774
MLP (W2V+Sick)	0.745	0.745	0.745
MLP (W2V+Object)	0.749	0.749	0.749
MLP (All)	**0.788**	**0.789**	**0.789**

表 4　主体性判定の結果

素性	適合率	再現率	F1
SVM (W2V)	0.673	0.672	0.672
SVM (W2V+ Pron)	0.690	0.690	0.690
SVM (W2V+Area)	0.685	0.683	0.683
SVM (W2V+Move)	0.676	0.676	0.676
SVM (All)	0.698	0.699	0.699
LR (W2V)	0.662	0.662	0.662
LR (W2V+ Pron)	0.674	0.674	0.674
LR (W2V+Area)	0.668	0.665	0.665
LR (W2V+Move)	**0.660**	0.661	0.661
LR (All)	0.679	0.679	0.679
MLP (W2V)	0.661	**0.660**	**0.660**
MLP (W2V+ Pron)	0.685	0.684	0.684
MLP (W2V+Area)	0.673	0.673	0.673
MLP (W2V+Move)	0.666	0.662	0.662
MLP (All)	**0.703**	**0.702**	**0.702**

表 4 より，最も適合率が高い組み合わせは，MLP に全ての検討素性を追加した場合で 0.788 を得た．これに対して，LR に baseline 素性である単語の分散表現と Move 素性を追加した場合に最も適合率が低く，0.660 となった．

Pron 素性及び，Area 素性を追加することで全ての分類器において，分類性能が 0.01 ポイント以上向上した．1 人称対代名詞に関する素性は，対象者を「当人」または「遠方者」に分類する際，重要であることを確認した．さらに，Area 素性を追加することで分類性能が向上したが，「当人又は近辺者」のラベルが付与されたツイートにおいて都道府県名が含まれる割合は全体の 1.1%，「遠方者」では 6%であった．このことから，都道府県名の有無が「当人又は近辺者」への分類に貢献したと考えられる．

4.4 エラー分析

症状の事実性解析では，短文ツイートに誤分類が多く見受けられた．症状が過去から現在に至るまで継続していると判断できるツイートは正例として扱うが，「僕，ずっーと頭痛い」のような短文ツイートの場合，症状の原因が病気によるものか否かの分類が困難である．短文ツイートについては，前後ツイートを利用することで改善できる可能性はある．また，症状の種類で分類が困難なツイートも確認できた．特に，「頭痛」，「めまい」の 2 症状は，様々な状況において言葉の比喩で用いられることが多く，分類が困難なものが存在することを確認した．

主体性の判定では，1 ツイート内に複数の人物が登場する場合，症状が発生している人物が誰であるかを判定することができず，誤分類している傾向を確認した．また，本稿では，30 種の一般的な 1 人称代名詞を辞書として利用したが，ツイート内には辞書に未登録の 1 人称代名詞も存在するため，特殊な代名詞が出現する場合において誤分類することを確認した．ゼロ代名詞のツイートも多いことから，照応解析の導入が必要である．

5 おわりに

本稿では，症状の事実性解析手法及び，ツイート文内の症状の主体性を判定する手法について検討した．3 種類の分類器に検討素性を組み合わせた実験により分類性能を評価した結果，症状の事実性解析では 0.788，主体性の判定では 0.703 という適合率を得た．今後は，さらに分類性能を向上させるための素性検討及び，症状以外の病名に対する事実性解析器の構築，病気・症状の動向をリアルタイムで可視化できるシステムの構築を目指す．

参考文献

[1] 北川善彬，小町守，荒牧英治，岡崎直観，石川博．インフルエンザ流行検出のための事実性解析．言語処理学会第 21 回年次大会（NLP2015）発表論文集，pp.218-221，2015.

[2] 松田紘伸，吉田稔，松本和幸，北研二．Twitter を用いた病気の事実性解析及び知識ベース構築．人工知能学会第 30 回全国大会（JSAI2016）論文集，pp.2C5-OS-21b-4，2016.

[3] Simin Ma, Xiaoquan Lai, Zhe Chen, Shenghao Tu, Kai Qin. Clinical characteristics of critically ill patients co-infected with SARS-CoV and the influenza virus in Wuhan, China. International Journal of Infectious Diseases, Vol.96, pp.683-687, 2020.

[4] AI Japan R&D Network. （オンライン）（引用日: 2021 年 1 月 13 日）https://www.ai-japan.go.jp/.

[5] 難波洋平， 安藤一秋．罹患者情報の可視化へ向けた Twitter の調査・分析．電気・電子・情報関係学会四国支部連合大会講演論文集 2017，p.226，2017.

[6] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomiki Ohkuma, and Eiji Aramaki. Overview of the NTCIR-13: MedWeb Task. In Proc. of the 13th NTCIR Conference, pp.40-49, 2017.

[7] 安藤翼，安藤一秋．ツイートされる病気症状の可視化に向けた病気症状の事実性判定手法の検討．FIT2019（情報科学技術フォーラム 2019）講演論文集，第 2 分冊，pp.139-140，2019.

[8] 松野省吾，水木栄，榊剛史．日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築．人工知能学会第 33 回全国大会（JSAI2019）論文集，pp.1-3，2019.

[9] 松吉俊，佐藤理史，宇津呂武仁．日本語機能表現書の編纂．自然言語処理，Vol.14，No.5，pp.123–146，2007.

[10] 株式会社メディカルノート 病名一覧．（オンライン）（引用日: 2021 年 1 月 13 日）https://medicalnote.jp/diseases/list.

[11] 日本語の一人称代名詞．（オンライン）（引用日: 2021 年 1 月 13 日）https://ja.wikipedia.org/wiki/日本語の一人称代名詞.