5N-02

ツイートされる多様な病気症状の可視化に向けた病気症状の事実性解析の検討

安藤　樹†　　安藤　一秋‡

香川大学大学院 創発科学研究科†　　香川大学 創造工学部‡

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が注目されている．たとえば，SNS を対象とした研究として，自身のツイート内容をもとに「うつ病」を診断する研究[1]や，インフルエンザに関するツイートを収集してインフルエンザの流行度合を推定する研究[2]，感染症のみを対象として病気の事実性を解析する研究[3]などがある．これらの研究は，対象を「特定の病気」に限定したものが中心である．しかし，SNS 上では，様々な病気や症状に関する内容が発信されており，これらの情報を活用することで新たな知見が得られる可能性がある．

そこで，当研究室では，特定の病気や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[4]の構築を進めている．このシステムにより，特定のエリアで特定の症状に関するツイートが急激に増加している場合，そのエリアで何か問題が発生していることが検知できる．新型コロナに感染した場合も様々な症状が発症していることが確認されており，症状に注目する意義はある．

先行研究[4]では，一般的な病気症状 14 種のみを事実性解析の対象に設定していたため，対象外となる病気症状が多数存在していた．そこで本稿では，病気症状の様々な患者表現を収集・分析している研究[5]で構築された患者表現辞書[6]に含まれる一般的な表現を利用して，病気症状 11 種に対する 86 表現を対象にした事実性解析手法について検討する．

2. 病気症状の事実性解析手法の検討

2.1. データセットの構築

本稿では，MEDNLP が公開している患者表現辞書[6]を利用して，患者表現辞書にある標準病名から 11 種の病気症状（標準病名）のいずれかを含むツイートを収集する．表 1 に対象とした標準病名を示す．患者表現辞書には，標準病名に対して，様々な出現系（患者がよく使う表現）が登録されている．たとえば，頭痛の場合，63 種の出現系が登録されている．API のリクエスト数の関係上，本調査では，出現系が 10 通りを超える場合，Google 検索のヒット数をもとに対象とする出現系を絞りこみ，各症状において最大 10 表現をキーワードに利用してツイートを収集する．

11 種の標準病名について，各 3,000 件ずつツイートを収集し，ツイート内に含まれる病気症状が病気に関係する場合は正例，病気によらない場合は負例として人手でラベルを付与する．本稿では，標準病名ごとに，収集した 3,000 件のツイートから正例と負例を 400 件ずつ抽出した計 800 件のツイートをデータセットとして利用する．

表 1　対象とした病気症状（標準病名）

頭痛	腹痛	胸痛	眼痛	耳痛	関節痛
咽頭痛	発熱	めまい	動悸	嘔吐感	

2.2. 事実性解析の素性検討

本稿では,先行研究[4]で使用している,つつじ素性（Ttj），病名素性，Zunda 素性（zd），単語分散表現（W2V）の 4 素性に，新しく時制素性を追加した 5 素性を用いる．なお，病名素性については，先行研究[4]で用いた病名の有無のみではなく，病名・症状の共起回数を利用した病名・症状素性（Sick）を利用する．以下に，本稿での実験で利用した素性を示す．以降，標準病名の出現系を病気症状と呼ぶ．

①つつじ素性（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．素性には，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，病気症状の右側最長 α 字中に含まれる最長またはすべての機能表現の意味 ID をそれぞれ Lm 素性，Morph 素性として利用する．

②Zunda 素性（zd）

Zunda は文中の「だろう」，「かもしれない」などのイベントに対して真偽判定や仮想性などを解析するモダリティ解析器である．Zunda の解析結果に含まれ真偽判定タグにおいて，各文に含まれる病気症状の後に続く動詞又はサ変接続名詞で一番近いイベントに付与されたラベルを素性とする．

③単語分散表現（W2V）

単語分散表現は，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[7]」を用いる．ツイートのベクトルには，文内の形態素に対する分散表現の平均を利用する．

④病名・症状素性（Sick）

ツイート内に病名または症状が共起する場合，病気に関係する可能性が高いといえる．そこで，文中に病名または症状の種類が共起する回数を素性として利用する．

Consideration of Factual Analysis of Disease Symptoms for Visualization of Various Disease Symptoms in Tweets

†Tatsuki Ando・Graduate School of Science for Creative Emergence, Kagawa University

‡Kazuaki Ando・Faculty of Engineering and Design, Kagawa University

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

⑤時制素性（Time）

ツイート発信時の状況だけでなく，過去や未来の内容が書かれている場合もある．そこで，ツイート内容が過去，現在，未来のうち，いつの内容であるかを素性として利用する．

3. 評価実験

3.1. 実験設定

本実験では，11 種類の病気症状に対し，二値分類問題で事実性を判定する．分類器としては，先行研究[4]と同様，Support Vector Machine（SVM），ロジスティック回帰（LR），多層パーセプトロン（MLP）を利用し，各素性を組み合わせることで，それぞれの判定性能を評価する．なお，SVM のカーネルは RBF，コストパラメータは 1,000，MLP は，最適化関数を SGD，バッチサイズを 128 とする．ツイート文の単語分割には，MeCab を利用する．

病気症状ごとに正例と負例が 400 ツイートずつ含まれるデータセットを利用する．適合率，再現率，F 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．本研究では，病気・症状を含むツイートの動向を可視化することを目的としているため，適合率を最も重視する．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．

3.2 実験結果

各病気症状に対して，分類器別に適合率が最良となった素性の組み合わせを表 2 に示す．表 2 に示す結果より，最も適合率の高い結果を得た病気症状は「めまい」であった．手法としては，MLP に W2V とつつじ素性，病名・症状素性，時制素性の 4 素性を組み合わせたもので，適合率 81.4%を得た．一方，最も適合率の低い結果となった病気症状は，「胸痛」であった．手法としては，MLP に W2V とつつじ素性，病名・症状素性の 3 素性を組み合わせたもので，適合率は 60.1%に留まった．

全体的によい結果を得た素性の組み合わせは，W2V とつつじ素性，病名・症状素性，時制素性を組み合わせたものであり，11 病気症状のうち，9 症状で適合率の最良値を得た．また，分類器も同様，，9 症状で MLP が適合率の最良値を得た．病気症状により，表現の種類数や負例となるツイートの特徴が大きく変わるため．分類性能にばらつきが生じていると考える．今後は，各症状のエラーについて分析する．

4. おわりに

本稿では,ツイートされる病気・症状を地域別・時系列別に可視化するシステムを実現するため，患者表現辞書に含まれる 11 種の病気症状を含むツイートを対象に，事実性解析する手法について検討した．

3 種類の分類器を用いた実験を通じて，各病気症状の分類結果を比較した結果，適合率の最良値は，「めまい」の 81.4%，最低値は「胸痛」の 60.1%であることを確認した．

今後は，分類性能を向上させるために，エラー分析すると共に，ニューラル言語モデルを利用した事実性解析手法について検討する．

表 2　判定結果

		precision	recall	F1
頭痛	SVM(W2V+Ttj+Sick+Time)	0.719	0.723	0.722
	LR(W2V+Ttj+Sick+Time)	0.711	0.719	0.718
	MLP(W2V+Ttj+Sick+Time)	**0.729**	**0.721**	**0.725**
腹痛	SVM(W2V+Ttj+Sick+Time)	0.712	0.706	0.710
	LR(W2V+Ttj+Sick+Time)	0.729	0.723	0.725
	MLP(W2V+Ttj+Sick+Time)	**0.731**	**0.742**	**0.735**
胸痛	SVM(W2V+Ttj+Sick)	0.598	0.585	0.592
	LR(W2V+Ttj+Sick)	0.593	0.602	0.594
	MLP(W2V+Ttj+Sick)	**0.601**	**0.613**	**0.608**
眼痛	SVM(W2V+Ttj+Sick+Time)	**0.624**	**0.629**	**0.625**
	LR(W2V+Ttj+Sick+Time)	0.598	0.602	0.603
	MLP(W2V+Ttj+Sick+Time)	0.613	0.604	0.608
耳痛	SVM(W2V+Ttj+Sick+Time)	0.588	0.602	0.593
	LR(W2V+Ttj+Sick+Time)	0.598	0.615	0.606
	MLP(W2V+Ttj+Sick+Time)	**0.611**	**0.617**	**0.613**
関節痛	SVM(W2V+Ttj+Sick+Time)	0.643	0.651	0.648
	LR(W2V+Ttj+Sick+Time)	0.651	0.650	0.650
	MLP(W2V+Ttj+Sick+Time)	**0.668**	**0.673**	**0.671**
咽頭痛	SVM(W2V+Ttj+Sick+Time)	0.674	0.686	0.681
	LR(W2V+Ttj+Sick+Time)	0.679	0.693	0.685
	MLP(W2V+Ttj+Sick+Time)	**0.697**	**0.714**	**0.706**
発熱	SVM(W2V+Ttj+Sick+Time)	0.698	0.721	0.713
	LR(W2V+Ttj+Sick+Time)	0.687	0.703	0.696
	MLP(W2V+Ttj+Sick+Time)	**0.719**	**0.722**	**0.719**
めまい	SVM(W2V+Ttj+Sick+Time)	0.779	0.788	0.783
	LR(W2V+Ttj+Sick+Time)	0.771	0.765	0.769
	MLP(W2V+Ttj+Sick+Time)	**0.814**	**0.798**	**0.809**
動悸	SVM(W2V+Ttj+Sick)	**0.682**	**0.647**	**0.662**
	LR(W2V+Ttj+Sick)	0.641	0.620	0.633
	MLP(W2V+Ttj+Sick)	0.667	0.675	0.67
嘔吐感	SVM(W2V+Ttj+Sick+Time)	0.592	0.613	0.604
	LR(W2V+Ttj+Sick+Time)	0.601	0.635	0.617
	MLP(W2V+Ttj+Sick+Time)	**0.621**	**0.635**	**0.628**

参考文献

[1] 玉井他，“うつ傾向推定に向けたツイート内容の解析法についての一検討”，言語処理学会第 22 回年次大会発表論文集，pp.385-388，2016.

[2] 北川他，“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 回年次大会発表論文集，pp.218-221，2015.

[3] 松田他，“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 回全国大会論文集，pp.2C5-OS-21b-4，2016.

[4] 安藤他，“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解析のための素性検討”，情報科学技術フォーラム講演論文集，pp.139-140，2019.

[5] 西谷他，“生成アプローチによる患者表現の標準化”，JAMI & JSAI AIM 合同研究会資料，pp.5-01-5-07，2021.

[6] 患者表現辞書，https://sociocom.naist.jp/patient-dic/

[7] 松野他，“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”，2019 年度人工知能学会全国大会（第 33 回）論文集，pp.1-3，2019.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.