

BART を用いたツイートされる多様な病気症状の事実性解析 Factuality Analysis of Various Disease Symptoms in Tweets using BART

安藤 樹[†] 安藤 一秋[‡]
Tatsuki Ando Kazuaki Ando

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。また、症状についても特定の病気にかかわるものしか対象としていない。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。新型コロナウイルス感染症の場合、幅広い症状が確認されており、症状の動向から未知の病気の発症状況を早期に検出できる可能性がある。

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[4-6]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。

提案システムを実現するため、病気・症状を含むツイートが投稿者の病気症状の発症状況について述べているか否かを判定する事実性解析[4,5]が必要となる。たとえば、“明日から期末試験が始まると思ったらめっちゃ頭痛いわ”や“さっきのつっこみ、おもしろすぎて腹痛い”といったツイートは、「頭痛い」と「腹が痛い」という症状を含むが病気症状ではない。さらに、地域別に可視化するためには、投稿者の居住地を推定する必要がある[6]。

先行研究[5]では、患者表現辞書[7]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対象に、多層パーセプトロンに単語分散表現[8]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モデルの分類性能が最良であることを確認した。

本稿では、分類性能を向上させるために、事前学習済みモデル BART と BERT を利用した事実性解析手法について検討する。なお、本稿では、BERT のみを用いたモデルの結果について述べる¹。

2. 先行研究

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としてツイートを収集していた。そのため、収集対象外となる病気・症状表現を含むツイートが多数存在

表 1 対象とした 11 種の病気症状（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	

していた。そこで、我々の先行研究[5]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[7]で構築された患者表現辞書[8]に含まれる一般的な表現を利用して、病気症状（標準病名）11 種に対する 86 表現を対象とする事実性解析手法について提案した。

図 1 に 11 種の病気症状を示す。先行研究[4,5]では、つつじ素性、Zunda 素性、単語分散表現、病名・症状素性、時制素性と 3 種類の分類器（Support Vector Machine, ロジスティック回帰、多層パーセプトロン）を組み合わせた手法を提案した。実験結果より、多層パーセプトロンに単語分散表現[9]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モデルの性能が 11 種類の病気症状のうち、9 種で最良を得ることを確認した。

本稿では、分類性能を向上させるため、先行研究[5]で対象にした病気症状 11 種 86 表現について、大規模事前学習済みモデル BERT を用いた事実性解析手法について、分類性能を確認する。

3. 提案手法

本手法では、事前学習モデルとして、Microsoft が提供している PubMedBERT の日本語バージョンをベースに利用する。PubMedBERT は、世界の主要な医学系雑誌に掲載された論文の書誌情報を調べることができるデータベースである PubMed の要約と PubMedCentral の全文記事を使用し事前学習されたモデルである。

本稿では、BERT の重み計算を最大限活かした予測ができるように PubMedBERT をベースとして、LSTM (Long Short-Term Memory) と次元畳み込み層 Conv1D、Linear 層を追加したモデルを構築する。また、ツイート内容が病気症状によるものを正例、病気症状によらないものを負例とした 2 種類のツイートでファインチューニングする。

4. 実験設定

本実験では、11 種の病気症状ごとに 5,000 件ずつツイートを収集し、そのうち正例と負例が 750 件ずつの計 1,500 件を学習データ、500 件をテストデータとして利用する。そして、11 種の病気症状ごとに事実性解析した結果で評価する。

また、学習データとして用いる各病気症状に関する 1,500 件のツイートに対し、ツイート内の文の順番をランダムに変更及び 2 文以上のツイートに対して病気症状を含まない文を削除することで、データを拡張した。

[†] 香川大学 大学院 創発科学研究科, Graduate School of Science for Creative Emergence, Kagawa University

[‡] 香川大学 創造工学部, Faculty of Engineering and Design, Kagawa University

表 2 各病気症状の F 値の平均

提案モデル	先行研究[5]のモデル
0.716	0.682

表 3 各病気症状の分類結果の F 値

	提案モデル	先行研究[5]のモデル
頭痛	0.769	0.725
腹痛	0.742	0.735
胸痛	0.688	0.608
眼痛	0.724	0.625
耳痛	0.688	0.613
関節痛	0.764	0.671
咽頭痛	0.712	0.706
発熱	0.732	0.719
めまい	0.786	0.809
動悸	0.629	0.662
嘔吐感	0.645	0.628

5. 実験結果

表 2 に、11 種の病気症状に対する F 値の平均値を示す。また、表 3 に、提案モデルと我々の先行研究[5]で提案したモデル（先行モデル）の病気症状別の最良の F 値を示す。

まず、表 2 に示す結果より、11 種の病気症状に対する F 値の平均値は、提案モデルで 0.716 となり、先行モデルの F 値 0.682 と比較して、3.4 ポイント向上した。しかし、本研究で構築を目指す地域別・時系列別の病気・症状可視化システムで利用することを考えた場合、F 値で 0.8 以上を目指したいと考えている。そのため、提案モデルの性能もまだ改善の余地があるといえる。

次に、表 3 の結果から、提案モデルにおいて、F 値が最良となった病気症状は「めまい」であり、0.786 を得た。また、提案モデルと先行モデルの両方で最良の F 値を得た病気症状は、どちらも「めまい」であるが、提案モデルの F 値の方が先行モデルよりも 2.3 ポイント低い結果となった。しかし、先行モデルにおいて、F 値が最低であった「胸痛」に対しては、提案モデルの方が 8 ポイント高い結果になった。「関節痛」に対しては、もっとも改善幅が大きく、提案モデルの F 値 0.764 は、先行モデルより 9.3 ポイント高い結果となった。

さらに、提案モデルと先行モデルの分類結果全体を比較すると、11 病気症状のうち、「めまい」と「動悸」以外の 9 種の病気症状に対して、提案モデルの分類性能が向上するという結果を得た。また、先行モデルにおいて F 値が 0.7 より低い病気症状については、提案モデルの方が、F 値が高くなりやすい傾向にあることが確認できた。

このような結果となる要因について考察する。「胸痛」や「耳痛」に関する負例には、感情表現として“胸が痛い”や“耳が痛い”と表現するツイートが多数存在する。提案モデルでは、先行モデルと比較して、BERT を用いることで前後の文脈をより考慮できるといえる。また、「眼痛」と「関節痛」については、“コンタクトを入れて目が痛い”といった、怪我でも病気症状でもなく単に「痛み」がある

ことを伝えるツイートや、や“瞼を切って「目が痛い」”といった、怪我による単なる「痛み」を伝えるツイートなど、負例のパターン数が他の症状より多様であるといえる。そのため、“コンタクトを入れて目が痛い”や“目が痛くて開かない。コンタクトのせいかな”といった、意図が同じ負例ツイートであっても、表現の多様性によって、先行モデルでは正しく分類できない場合があった。文脈性が考慮できる提案モデルにおいては、このような問題を軽減できたと考えられる。

6. おわりに

本稿では、病気・症状の事実性解析の性能向上を目指し、大規模事前学習済みモデル BERT を用いた病気・症状の事実性解析手法について検討した。先行研究[5]で提案した、つつじ素性、Zunda 素性、単語分散表現、病名・症状素性、時制素性と 3 種類の分類器を用いたモデルの性能よりも、F 値で 3.4 ポイント高い性能を得ることを確認した。しかし、「めまい」と「動悸」に対する分類性能は、先行研究で提案したモデルよりも、それぞれ 2.3 ポイント、3.3 ポイント低い結果となった。提案モデルにおいて、11 種の病気症状の F 値の平均値は、0.716 にとどまり、事実性解析の性能については改善の余地があるといえる。

今後は、より分類性能を向上するために、ファインチューニング手法や最適なパラメータ値などを模索するとともに、他のニューラルネットワークモデルや事前学習済みモデルなどを用いて、病気・症状を事実性解析する手法について検討する。

参考文献

- [1] 玉井森彦, 米澤(深谷)拓吾, 川西直, 長谷川晃朗, 武内良男, 田近亜蘭, 小川雄右, 古川壽亮, “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 回年次大会発表論文集, pp.218-221, 2015.
- [3] 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 回全国大会論文集, pp.2C5-OS-21b-4, 2016.
- [4] 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] 安藤樹, 安藤一秋, “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”, , 2023.
- [6] Masahiro Matsumoto, Kazuaki Ando, “A Deep Learning Model of Estimating User’s Place of Residence Using Tweets and Weather Information”, Proceedings of 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, 6 pages, Dec.18-20, 2022.
- [7] 西谷実紘, 矢田竣太郎, 若宮翔子, 荒牧英治, “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.
- [8] 患者表現辞書, <https://sociocom.naist.jp/patient-dic/>
- [9] 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”, 2019 年度人工知能学会全国大会 (第 33 回) 論文集, pp.1-3, 2019.

¹ タイトルでは BART を利用すると記載しており、学会発表申込時は BART の結果を中心に BERT、先行研究で提案したモデルとの比較についてまとめる予定であった。しかし、原稿執筆時に BART の実験環境に問題がある可能性を発見したため、本稿では、BERT による結果のみを述べる。