

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 Factuality Analysis of Disease Symptoms for Visualization of Disease Symptoms based on Twitter

安藤 翼†
Tsubasa Ando

安藤 一秋‡
Kazuaki Ando

1. はじめに

近年、感染症を中心とする病気・医療分野に自然言語処理を応用する研究[1,2]が行われている。これらの研究の中心は、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、病気の事実性解析を行う研究[2]であり、他の病気・症状を対象とした研究はあまり行われていない。

そこで本研究では、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[3]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手法だけでなく、病名を含まないツイートに対する病気症状の事実性解析手法が必要となる。NTCIR-13 MedWeb Task[4]において、8つの病名又は症状を含むツイート文に対して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、インフルエンザなどの感染症であるか否かを問わず、Twitter に投稿されたツイートに含まれる様々な病気症状に対して、二値分類問題として事実性を解析する手法について検討する。

2. 病気症状に対する事実性解析手法の検討

2.1 データセットの構築

ツイートに含まれる病気症状に対して事実性解析する手法を検討するに先立ち、病気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、対象とする病気症状の辞書が必要となる。本稿で対象とする 14 病気症状を表 1 に示す。病気症状は性別、年齢に関係しない一般的なものとする。

表 1 対象とする病気症状

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、ツイート内に病気症状を必ず 1 語以上含むこと、及び、病気症状は年齢、性別は問わないこととする。

† 香川大学大学院工学研究科 Graduate School of Engineering, Kagawa University

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

以上の条件で収集した結果、合計 1,471 ツイートが得られた。収集したツイートを文単位で分割し、病気症状が含まれる「文（ツイート文）」に対して、病気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付与する。

以下、ラベル付与の 4 つの基準について示す。

- ① 対象者は本人及びその近辺者
ツイート文の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。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
ツイート文から現在起きていると判断できる場合は「真」、過去に起きた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。
- 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
ツイート文内の病気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。
なお、病気症状が含まれるツイート文のうち、1 文では真偽判定できない場合は利用しない。

以上の条件で真偽ラベルを付与した結果、正例は合計 735 件、負例は合計 736 件となり、1,471 件のラベル付きツイート文からなるデータセットを構築した。

2.2 事実解析の素性検討

本稿では、病気症状の事実性を二値で判定するため、以下の 5 つの素性について検討する。

- ① Bag of Words (BoW)
データセット内の各ツイート文には、表 1 の病気症状が必ず 1 語以上含まれる。そのため、文内には、病気症状に関する特徴的な語が存在する可能性が高い。そこで、単語の出現頻度を考慮した Bag of Words によりベクトル化し、それを素性候補とする。
- ② 病名の有無 (Sick)
文中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候補とする。病名は「株式会社メディカルノート[5]」に登録されている 2,102 種の病名を辞書利用する。
- ③ 体の部位の有無 (Parts)
症状が病気と関係する場合、その症状が発生している体の部位が文中に含まれる可能性が考えられる。そこ

で体の部位の有無も素性候補とする。体の部位は、厚生労働省が公開する「人体各器官図[6]」を基に、体の 48 部位を辞書利用する。

④ 文長 (Len)

病気症状を含む各文の文字数を素性とする。データセット内の各文の文字数が、正例文の場合は平均約 26.0 文字、負例文の場合は約 38.8 文字であり、平均文字数に 12.8 文字の差があるため、素性候補とする。素性は、各文の文字数をデータセット内の最大文字数で割り、正規化した値を用いる。

⑤ 単語分散表現 (W2V)

単語分散表現を素性候補とする。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。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。素性には、文内の形態素に対する分散表現の平均を利用する。

3. 評価実験

検討素性の有効性を確認するために、評価実験を行う。

3.1 実験方法

本実験において、形態素解析には MeCab を利用する。実験データには、2.1 で構築したデータセット（正例：735 件；負例：736 件）を用いる。

実験には、Support Vector Machine (SVM)、ロジスティック回帰 (LR)、多層パーセプトロン (MLP) の 3 種類の分類器を用いる。実装には scikit-learn を使用し、分類器ごとに 5 つの検討素性を組み合わせ、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。

3.2 評価方法

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。本実験では、事実に解析の結果、ツイート文に含まれる病気症状が病気に関係するものか否かの二値で分類する。そのため、適合率を最も重視し、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。

3.3 実験結果

3 種類の分類器に対して、5 種類の検討素性の組み合わせによる実験結果を表 2 に示す。実験結果より、最も適合率が高かったのは、SVM において分散表現を素性とした場合で、0.743 という結果が得られた。これに対して、最も適合率が低いものが、MLP において Bag of Words を素性とした場合で、0.623 であった。

Bag of Words を素性としたときは、SVM、LR、MLP 共に適合率が 0.7 未満となり、他の素性との組み合わせと比較して、分類性能が低いことを確認した。

病気の Sick 素性及び、体の部位の Parts 素性、文長の Len 素性はいずれの分類器においても適合率が低下した。原因として、Sick 素性、Parts 素性は構築したデータセット内に含まれる数が少ないことが挙げられる。データセット内に Sick 素性が含まれるのは 123 文であり、全体の約 8%、Parts 素性は 290 文で全体の約 19%しか含まれて

いないため、素性として有用でなく、分類性能の低下を招いたと考えられる。

構築したデータセットは、病気症状が含まれるツイート文に対して、正解ラベルを付与したものである。ツイート文には、その前後の文あるいは前後のツイートが活用できる。これらの情報および新たな素性を検討することで、さらに分類性能を向上できる可能性が考えられる。

表 2 分類結果

	precision	recall	f1
SVM+BoW	0.642	0.637	0.638
SVM+W2V	0.743	0.735	0.735
SVM+W2V+Sick	0.742	0.734	0.734
SVM+W2V+Parts	0.738	0.734	0.733
SVM+W2V+Len	0.741	0.735	0.735
LR+BoW	0.690	0.683	0.684
LR+W2V	0.722	0.716	0.716
LR+W2V+Sick	0.707	0.700	0.702
LR+W2V+Parts	0.704	0.698	0.700
LR+W2V+Len	0.707	0.702	0.703
MLP+BoW	0.623	0.690	0.690
MLP+W2V	0.727	0.725	0.724
MLP+W2V+Sick	0.726	0.732	0.720
MLP+W2V+Parts	0.724	0.717	0.717
MLP+W2V+Len	0.724	0.714	0.714

4. おわりに

本稿では、ツイートされる病気症状の地域別・時系列別に可視化を実現するため、様々な病気症状に対する事実に解析手法および素性について検討した。3 つの分類器及び 5 つの素性により分類性能を評価した結果、SVM において分散表現を素性とする場合、最も分類性能が高く、0.743 という適合率を得た。

今後は、ツイート文の前後文や前後のツイートの利用と新しい素性について検討する。

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に解析”, 言語処理学会第 21 回年次大会(NLP2015), 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に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), pp.2C5-OS-21b-4, (2016).
- [3] 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”, 電気関係学会 四国支部連合大会 2017(IEICE2017), p.226, (2017).
- [4] Wakamiya, S., Morita, M., Kano, Y., Ohkuma, T. and Aramaki, E., “Overview of the NTCIR-13: MedWeb Task”, In Proceedings of the 13th NTCIR Conference, pp.40-49, (2017).
- [5] 株式会社メディカルノート 病名一覧 (2018/12/17 閲覧) <https://medicalnote.jp/diseases/list>
- [6] 厚生労働省 人体各器官図 (2019/06/21 閲覧) <https://www.mhlw.go.jp/file/06-Seisakujouhou-10800000-Iseikyoku/0000209868.pdf>
- [7] 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), pp.1-3, (2019).