Twitter上の対話に着目した罹患者への定型的応答の自動抽出

浅川 玲音　　秋葉 友良
豊橋技術科学大学 情報・知能工学専攻
{asakawa.reine.ci, akiba.tomoyoshi.tk}@tut.jp

1 はじめに

疾病の発生動向を監視し迅速かつ的確な対策をとる疾病サーベイランスは，多様な疾病の発生と蔓延を防止する為に重要な課題である[1]．従来は病院等が収集した情報をもとに行われてきたが，近年ではTwitterベースの疾病サーベイランスシステムが提案されている．Twitterは一般の多くの人々が発信するリアルタイムの情報を大量に保有しているという点が疾病サーベイランスシステムに適していると考えられ，Twitterベースの疾病サーベイランスシステムは世界中で研究され始めている [1]．Twitterベースのアプローチでは，まず対象疾患の名前を含むツイートを収集し，その中から特に疾患への罹患を話題にしたツイートを検出し集計することで流行を予測する．ツイートの罹患判定の先行研究の多くは教師ありの機械学習を利用しており，いずれもある程度の学習コーパスを人手でアノテーションして作成する必要がある [2]．しかしながら，人手によるラベリングはコストが高く，病気・症状への罹患を表す多様な表現をカバーするツイートのサンプルを用意することが困難であるといった理由から，学習コーパスを十分に用意することは非常に困難であると言える．

ツイートの罹患判定器の学習データ不足の問題に対し，「お大事に」をクエリに自動的に学習コーパスを獲得する手法が提案されている [3]．罹患者に対して「お大事に」と声をかける日常会話に着目し，「お大事に」を含むツイートにリプライされているツイート，見做し罹患ツイート，を罹患Positiveなツイートの学習データとして利用することで罹患判定器の学習データを拡張する手法である．しかしながら，浅川らの手法では「お大事に」とリプライされ得る罹患Positiveなツイートしか取得することができないという問題がある．そのような問題に対し，本稿では「お大事に」を含む対話の構造に着目して罹患者への定型的応答を自動的に獲得する手法を検討した．見做し罹患ツイートのリプライには「お大事に」以外にも「養生して下さい」等の特に罹患者に対してかけられやすい言葉が含まれている．見做し罹患ツイートのリプライに頻出する文字列パターンを抽出し，罹患者への定型的応答の信頼性スコアを計算した．[3] で提案された罹患判定器を用いて信頼性スコアの妥当性を検討し，罹患判定器の学習データ自動獲得法のクエリ拡張の可能性を示唆した．本稿の構成は次の通りである．まず第2章で関連研究について述べ，第3章で提案法の詳細を紹介し第4章で実際に獲得した罹患者への定型的応答候補と信頼性スコアについて述べ，最後に第5章で結論と今後の課題をまとめた．

2 関連研究

大量の文書の中から何らかの手がかり表現を用いて目的の情報を取得する研究は多くの分野で取り組まれてきた．特許明細書からの情報抽出の分野では，「ことにより」という手がかり表現を利用して特許明細書から手段とその効果で構成される因果関係知識を抽出する手法が石川らによって研究されている [4]．「ことにより」は有力な手がかり表現であるが，「ことにより」を使用していない文からの因果関係知識を抽出することができないという問題が指摘されている．

1つの手がかり表現では情報を取得しきれないという問題に対して，「ができる」という1つの手がかり表現から大量の手がかり表現を自動的に獲得する手法が酒井らによって提案されている [5]．酒井らの研究は膨大な特許出願情報を可視化するパテントマップの自動生成のために，特許明細書から技術課題及び解決手段を抽出することを目的としている．最初に1つの手がかり表現「ができる」を人手で与え，手がかり表現の直前に出現する共通頻出表現候補を抽出し，様々

[1]厚生労働省，”感染症発生動向調査について”，https:www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000115283.html，参照 (2019/8/29).

な手がかり表現に修飾している共通頻出表現は適切であるという仮定のもとで共通頻出表現を選別する．次に共通頻出表現が係っている手がかり表現候補を抽出し，様々な共通頻出表現に修飾されている手がかり表現は適切であるという仮定のもとで新たな手がかり表現を選別する．この手順を新たな表現が得られなくなるか予め定めた回数まで繰り返す．候補から適切な表現を選別するための評価には前述の仮定のもとに式 1 が提案されており，その値が閾値以上の表現を選択することが提案されている．ここで，e は選別したい表現，s は確定している表現を表し，$S(e)$ は e と共起する s の集合を表している．$P(e,s)$ を求める式は，後に酒井と坂地らが提案した Cross-bootstrapping 法 [6] で採用されている式 2 を本研究では参考にした．

$$H(e) = -\Sigma_{s \in S(e)} P(e,s) \log_2 P(e,s) \quad (1)$$

$$P(e,s) = \frac{e \text{ と } s \text{ の共起数}}{e \text{ の数}} \quad (2)$$

本稿では，見做し罹患ツイートの自動獲得手法を応用して罹患者への定型的応答候補を抽出し，多様な病名・症状名を含むツイートを取得できた候補は適切であるという仮定のもとに，罹患者への定型的応答を選別することを試みた．本研究は酒井らの手法と次の点で異なっている．(1) 共通頻出表現を選別する代わりに一般的な病名・症状名のリストを用いている．これは様々な病気・症状への罹患を表すツイートを獲得できるクエリを獲得するのが本手法の目的であるので，酒井らのように共通頻出表現を自動獲得する必要がないからである．(2) 手がかり表現候補を見做し罹患ツイートのリプライから頻出パターンマイニングによって抽出する．これは罹患を話題としたツイートへのリプライに定型的な言葉を候補とするためである．

3 提案法

3.1 概要

日常生活において罹患者にかけられる言葉は「お大事に」だけではない．見做し罹患ツイートに対する「お大事に」以外のリプライを観察したところ，相手の症状を繰り返す・励ます・回復を願う・休息や治療法を提案する・原因について考察する・心配や驚きを表すといったリプライが多く見られ，「元気になって」「早く良くなりますように」「養生して」「薬を飲んで」といった定型的な言葉が見られることが分かった．本稿では以上の洞察から，見做し罹患ツイートへのリプライに頻出する文字列を罹患者への定型的応答候補として抽出した．

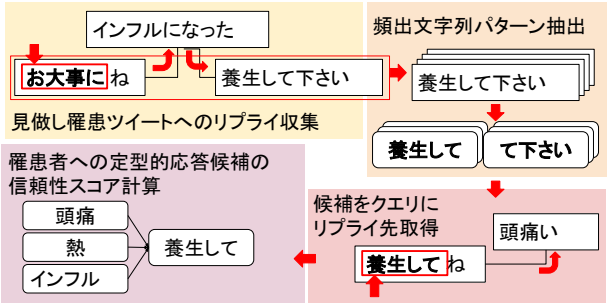

図 1: 罹患者への定型的応答の自動獲得

提案する罹患者への定型的応答の自動獲得法の全体像を図 1 に示す．まず最初に「お大事に」をクエリとして見做し罹患ツイートとそのリプライを収集し，リプライをまとめた文書から頻出文字列パターンを抽出する．抽出したパターンを罹患者への定型的応答の候補とし，各候補をクエリにツイートとそのリプライ先を収集する．最後に各候補の信頼性スコアを計算する．候補の抽出手順を 3.2 節に，信頼性スコアの計算方法を 3.3 節にまとめた．

3.2 文字列パターンマイニングを用いた罹患者への定型的応答候補の抽出

本節では罹患者へのリプライをまとめた文書から罹患者への定型的応答候補を抽出する手順についてまとめる．まず初めに収集した見做し罹患ツイートのリプライの本文を 1 つの文書にまとめ，語彙制限等の前処理を行った後，頻出文字列パターンマイニングを行うことで取得した．その後，抽出したパターンからツイートの検索クエリとしてふさわしくないパターンを削除し，包含関係にありそうなパターンの削除を行った．具体的な手順を下記に示す．

1. リプライ本文を 1 つの文書にまとめた
2. 正規表現を用いて“数，URL，ハッシュタグ，改行コード，文末”をタグ化した
3. MeCab を用いて分かち書きし，語彙と各単語の出現頻度を調査した
4. 出現頻度が 2000 より少ない単語は語彙から削除し，見做し罹患ツイートへのリプライ中の未知語を UNK タグ で置換した
5. 文字分割で Suffix Array を作成した[2]
6. Suffix Array を用いて長さ 4 から 20 までのパターンを抽出した

[2]Suffix Array の作成には，効率的なアルゴリズム「SA-IS」を C 言語で実装した「sais-lite」の Python ラッパー版「pysais-utf8」を利用した．https://github.com/lars76/pysais-utf8

7. 頻度 1000 以下のパターンを削除した
8. Twitter の検索クエリとして使用できないパターンを削除した
   - タグを含むパターンは削除．ただしパターンの末尾に文末タグがあるものはタグだけを除去して採用した
   - (,), (,) を含むパターンは削除した
9. 漢字を 1 文字も含んでいないパターンは削除した
10. 長さ M と長さ M+k の 2 つのパターンを比較したとき，それぞれの出現頻度が誤差 10%未満で一致し，かつ，先頭もしくは末尾から M 文字が完全一致するようなら，長さ M のパターンは削除した．

3.3 エントロピーを用いた罹患者への定型的応答の信頼性スコアの計算

本節では前節で抽出した候補から適切な罹患者への定型的応答を選別するための信頼性スコアについてまとめる．各候補が自動獲得法のクエリに適しているかどうかは，候補を用いて実際にツイートを収集し罹患クラスを評価することができれば確認できるが，その評価を人の手で行うのは非常にコストがかかり現実的ではない．そこで本研究では，多様な病名・症状名を含むツイートを自動獲得することができる候補が罹患者への定型的応答であると仮定して，各候補の信頼性スコアを計算する．罹患者への定型的応答候補は見做し罹患ツイートへのリプライに頻出した文字列パターンであり，罹患を話題としたツイートへのリプライに定型的に含まれる言葉であるといえる．そのような言葉の中で，様々な病名・症状名を含むツイートのリプライに用いられやすい言葉は，何らかの病気・症状への罹患を表すツイートを自動獲得するためのクエリとして相応しいのではないかという直感に基づいている．

一般的によく知られている病気や症状を共通頻出表現として利用することを考え，NHK 健康チャンネルのウェブサイトにおける「病名・症状から探す」のページ[3]に掲載されている病名と症状名を参考に病名・症状名のリストを作成した．病名は「50 音から病気を探す」に一覧されている 181 個の病名を利用し，インフルエンザなどの幾つかの病名は省略形や漢字・かなで表した形などの表記揺れへの対策をとった．症状名は「症状から病気を探す」に一覧されている症状を参考に，できるだけ名詞のみの形で症状を検索できるように 53 個の症状名を構成した．

[3]日本放送協会, "病名・症状から探す", https://www.nhk.or.jp/kenko/disease/, 参照（2019/8/29)

まず最初に，前節で抽出した罹患者への定型的応答候補をクエリにしてツイートとそのリプライ先のツイートを収集した．ここで罹患者への定型的応答候補を含むツイートによってリプライされているツイートのことを見做し罹患ツイート候補と称することにする．収集した罹患者への定型的応答候補を含むツイートとそのリプライ先の見做し罹患ツイート候補のペアから，罹患者への定型的応答候補の信頼性スコアを計算した．罹患者への定型的応答候補の集合を $E = e_1, e_2, ..., e_N$(N;候補数)，病名・症状名の集合を $S = s_1, s_2, ..., s_{234}$，$N_{s_j,e_i}$ を s_j を含む見做し罹患ツイート候補と e_i を含むツイートのペアの数とし，N_{e_i} を e_i を含むツイートの数とおいたとき，任意の罹患者への定型的応答候補 e_i の信頼性スコア h_{e_i} を式 3 により計算した．

$$h_{e_i} = -\Sigma_{s_j \in S} P_{e_i}(s_j) \log_2 P_{e_i}(s_j) \quad (3)$$

$$P_{e_i}(s_j) = \frac{N_{s_j,e_i}}{N_{e_i}} \quad (4)$$

4 実験と考察

実際に罹患者への定型的応答候補の自動抽出と信頼性スコアの計算を行い，信頼性スコアの妥当性を検討した．

候補の抽出に利用した見做し罹患ツイートとそのリプライのサイズと収集期間を表 1 に示す．3.2 節の手順により 803 個の罹患者への定型的応答候補を得た．出現頻度上位 28 個のパターンを表 2 に示す．表 2 を見ると,「ゆっくり休んで」「大丈夫ですか?」等の罹患者にかけそうな言葉が抽出できていることが分かる．一方で「下さいね」「良かった」等の罹患者にかける言葉としては想起しなさそうな言葉も抽出されている．また，例「大丈夫ですか？」は一見して罹患者への声掛けらしい言葉だが，異なるシーンでは非罹患者に対しても使用されることが考えられる．自動獲得法のクエリとして利用することを考えると罹患者に対してだけ使われるパターンを選別することが望ましい．

罹患者への定型的応答候補を含むツイートを各候補ごとに 200 件ずつ取得し，その見做し罹患ツイート候補を収集した．リプライ先を持っていないツイートが

表 1: 見做し罹患ツイートとそのリプライの収集期間

	見做し罹患ツイート	リプライ
ツイート数	257,151	1,180,551
keyword	お大事に	
収集期間	2019 年 1 月下旬～2019 年 2 月下旬（合計 12 日間）	

表 2: 罹患者への定型的応答の候補

ゆっくり休	り休んで	て下さい	ゆっくり休んで
大丈夫で	お疲れ様	大丈夫ですか	お疲れ様で
良かった	頑張って	お疲れ様でした	休んでください
気をつけ	大丈夫?	大丈夫ですか?	り休んでください
して下さい	下さいね	ゆっくり休んでください	気をつけて
思います	無理しない	と思いま	楽しみに
って下さい	と思います	て下さいね	無理しないで

あることや 1 つのツイート中に複数のクエリが含まれている場合があることにより，実際に取得された見做し罹患ツイート候補は 51,974 件となった．罹患者への定型的応答候補を含むツイートと見做し罹患ツイート候補のペアをもとにして，803 個全ての罹患者への定型的応答候補の信頼性スコアを計算した．比較のために「お大事に」の信頼性スコアも計算した．スコア上位 30 個と下位 30 個の罹患者への定型的応答候補を，表 3 に示す．表 3 を見ると，信頼性スコア上位の候補には罹患者に対してのみよく使われそうな表現が，信頼性スコア下位の候補には非罹患者に対してもよく使われそうな表現が挙げられており，直感的にも罹患者への定型的応答らしさをスコア化できているように思われる．

信頼性スコアの妥当性を検討するために，見做し罹患ツイート候補の罹患クラスを浅川ら [3] の罹患判定器を用いて推定し，判定結果と信頼性スコアとの相関を求めた．罹患判定器には浅川らが提案した 2STEP モデルにハイパーパラメータの変更や開発データの導入を加えたもの（論文の 2STEP モデルの Accuracy と比較すると NTCIR テスト −0.030(0.848)，実テスト +0.12(0.712)）を用いた．まず全ての見做し罹患ツイート候補に対して罹患判定を行ってから，各罹患者への定型的応答候補についてペアの見做し罹患ツイート候補の罹患判定結果の平均を求めることで，罹患スコアの列 $X = x_{e_1}, x_{e_2}, ..., x_{e_{803}}$ を得た．続いて信頼性スコアの列 $H = h_{e_1}, h_{e_2}, ..., h_{e_{803}}$ と罹患スコアの列 R のピアソンの相関係数を計算したところ，信頼性スコアと罹患スコアの相関係数は $r \sim 0.801$ となった．罹患スコアとの間に強い正の相関が確認されたことにより，自動獲得法のクエリとして有用な罹患者への定型的応答を選別するためのスコアとして，提案する信頼性スコアの妥当性が示唆された．

5 おわりに

見做し罹患ツイートの自動獲得法のクエリを拡張するために、罹患者への定型的応答を自動獲得する手法を提案した。実際に抽出した罹患者への定型的応答候補の信頼性スコアを計算し、罹患判定器 [3] による罹患判定結果と信頼性スコアとの間に強い正の相関を確認し，罹患判定器の学習データ自動獲得法のクエリ拡張の可能性を示唆した．今後の課題として，手順中の変数や信頼性スコアの閾値の妥当な値の範囲を決めるための基準について検討する必要がある．それから，新しい罹患者への定型的応答が自動獲得法のクエリとして利用可能かどうかを罹患判定器の分類精度を比較して検証する必要がある．また，新たな罹患者への定型的応答をもとにして見做し罹患ツイートとそのリプライを収集することで，酒井らの手法のようにブートストラップ的にクエリ拡張を行うよう手法を拡張することも考えられ，その場合には適切な終了条件の設定の検討などが課題となると考えられる．

表 3: 罹患者への定型的応答の候補

上位	罹患者への定型的応答候補	スコア	下位	罹患者への定型的応答候補	スコア
1	良くなりますように.	4.501	803	池江さん	0.000
2	良くなりますように	4.193	802	池江選手	0.000
3	早く良くなり	4.191	801	青木アナ	0.240
4	お大事に	4.189	800	で待って	0.285
5	早く良くなりますように	4.184	799	日頑張りま	0.287
6	良くなります	4.170	798	待ってます.	0.453
7	早く良くな	4.128	797	せて頂き	0.478
8	早くよくなりますように	4.074	796	て頂きま	0.496
9	ちゃん，大	4.043	795	せて頂きま	0.506
10	早く良くなる	3.882	794	了解です！	0.507
11	やん，大丈夫	3.872	793	応援してます.	0.528
12	早く良くなりますように.	3.802	792	慎吾ちゃん	0.536
13	早く良くなると	3.799	791	待っています.	0.544
14	辛いですよね	3.761	790	大丈夫ですよ	0.590
15	病院に行	3.752	789	頑張って！	0.613
16	早くよくな	3.728	788	と信じて	0.615
17	良くなる	3.700	787	楽しみに待ってます	0.626
18	早く治ります	3.686	786	が大好きで	0.630
19	良くなりますように！	3.676	785	に待ってます	0.631
20	治りますように	3.667	784	も大丈夫	0.637
21	良くなると	3.666	783	を応援し	0.649
22	早く良くなるといい	3.647	782	了解です	0.655
23	病院に行って	3.642	781	誕生日おめでとうございます！	0.660
24	早く良くなって	3.642	780	可愛すぎ	0.669
25	一日も早	3.611	779	応援します	0.670
26	早く治りますように	3.583	778	るの楽しみ	0.679
27	治ります	3.578	777	で大丈夫	0.699
28	ゆっくり休んで早く	3.564	776	好きです.	0.703
29	日も早く	3.554	775	のを楽しみにして	0.720
30	良くなるといい	3.551	774	の楽しみ	0.746

謝辞 本研究は JSPS 科研費 19K11980 の助成を受けた．

参考文献

[1] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proc. Conf. on EMNLP*, EMNLP '11, pp. 1568–1576, 2011.

[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. 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 2C5OS21b4–2C5OS21b4, 2016.

[3] 浅川玲音, 秋葉友良. 罹患者への定型的応答を利用した罹患ツイートの自動獲得と rnn 罹患判定器学習への適用. 言語処理学会大会発表論文集, Vol. 25, pp. 1177–1180, March 2019.

[4] 石川大介, 石塚英弘, 宇陀則彦, 藤原譲. 特許文献における因果関係の抽出と統合. 情報知識学会誌, Vol. 14, No. 4, pp. 105–118, 2004.

[5] 酒井浩之, 野中尋史, 増山繁. 特許明細書からの技術課題情報の抽出. 人工知能学会論文誌, Vol. 24, No. 6, pp. 531–540, 2009.

[6] 坂地泰紀, 野中尋史, 酒井浩之, 増山繁. Cross-bootstrapping：特許文書からの課題・効果表現対の自動抽出手法. 電子情報通信学会論文誌 D, Vol. 93, No. 6, pp. 742–755, 2010.