2C5-OS-21b-4

Twitterを用いた病気の事実性解析及び知識ベース構築

Facticity analysis of tweets about illness for symptom database construction

松田 紘伸　吉田 稔　松本 和幸　北 研二
Matsuda Hironobu　Yoshida Minoru　Matumoto Kazuyuki　Kita Kenji

徳島大学 工学部
Faculty of Engineering, Tokushima University

In recent years, enhancement of medical efficiency using information technology is popular. It is necessary for a general user to easily use such a technology. A purpose of this research is to make a database that predict illness by using Twitter. As a method for making the database, we judge a user infected with illness from tweets and get a symptom of illness from such users. As a result, we obtained a database for predicting illness.

1. はじめに

現在,医療・福祉分野の問題として,病気の早期発見を行う技術が求められている.日本人の病気での死亡数は,統計より,全死因の約 6 割にも及んでおり,問題は深刻化している.このような問題を解決するには,病気の早期発見を行うことが重要であると考えられている.病気の早期発見を行うことで,速やかに適切な治療を行うことができ,症状の悪化を防ぐことが可能となり死亡者数や,病気によって起こる後遺症の減少にもつながる.このような早期発見の技術を発展させることで,医療・福祉分野全体の発展も望める.

本研究では,社会的な要素を備えたコミュニケーションネットワークである Twitter を用いて情報収集を行う. Twitter はリアルタイム性に優れており,利用者数も多いことから多数の情報を効率的に取得するのに優れている.その情報を用いて現在,需要が高いと思われる医療・福祉分野への応用が可能だと考えた.この Twitter 上に投稿されるツイートの文章から病気の兆候・症状に関するデータを取得し,病気の早期発見を行う手法を提案する.

本研究では,Twitter に投稿されたツイートから病気の名称が含まれているツイートを抽出する.そのツイートの文章の内容から,投稿した人物が病気に感染しているか否かの事実性解析を行う.

事実性解析の結果,病気に感染していると判断された場合,感染前のツイートより,その病気の兆候・症状を取得する.その病気の兆候・症状をデータベースとして構築することを目指す.

2. 既存研究

既存研究としては,北川ら[1]は Twitter を対象とし,インフルエンザに感染している人物を,モダリティという当人の判断や感じ方を用いて事実性の解析を行い,その結果よりインフルエンザの流行検出を行うという手法を提案している.しかし,この研究ではインフルエンザのみを対象としており,他の病気への応用は未知数である.

また,既存研究では最終的な目的をインフルエンザの流行検出にしているが,本研究ではツイートから病気を予測するための知識ベースの構築を最終的な目的にしているため,多数の病気に関するデータが必要となる.

2.1 事実性解析

事実性解析とは,文章中の事象が実際に起こったか,起こらなかったかの事実性の判定を行う解析法である.

本研究では,事実性解析をツイート文章中に存在する病気を対象に行い,その病気に感染しているか,感染していないかの判定を行う.この事実性解析の手法として,SVM を用いた機械学習と,Zunda 解析器を用いたモダリティによる解析を用いて行う.

Zunda 解析器[2]とは,オープンソースの日本語拡張モダリティ解析器である. モダリティとは,話題に対する当人の判断や感じ方を表す言語表現である.この Zunda 解析器を用いて,文章中のイベントである動詞や形容詞に対して,そのイベントが起こったかどうかの真偽判断等を行う.

2.2 SVM

SVM(サポートベクターマシン)[3]とは,教師あり学習を用いて行うパターン認識の手法の 1 つである.SVM は,線形入力素子を利用して 2 クラスのパターン識別器を構成する手法である.

線形分離超平面上に,正例と負例の 2 つのクラスが存在し,この 2 クラスを分類する識別面を求めるとする.識別面が正例と負例,それぞれの最も近い事象との距離であるマージンが最大になることで,2 クラスを分類する識別面が求められる.この識別面を用いることで,未知の事象に対してのクラスを求めることができる.

本研究では,病気に感染しているツイートを”真”,病気に感染していないツイートを”偽”と定義し,予め真偽判断を行ったツイートを学習データとして,新たに取得したツイートの分類を行う.

3. 提案手法

本研究で提案する手法は,Twitter 上に投稿されているツイートの文章を対象とし,事実性解析を用いた真偽判断を行い,病気に感染している人物を判断する.その結果を用いて,病気に感染

連絡先:松田紘伸,c501406905@tokushimau.ac.jp
吉田稔,松本和幸,北研二,{mino,matumoto,kita}@is.tokushima-u.ac.jp

していると判断された人物より,病気の兆候・症状を検出する解析データを用いたデータ収集を行う.

Twitter 上に投稿されている病気の名称が含まれたツイートを入力として用い,そのツイートに事実性解析を行い,病気に感染している人物を判断する.そのツイートを投稿したユーザの過去のツイートを取得し,文章中に含まれる語を出力とすることで,その語を病気に感染する前の兆候・症状だと考える.最終的に評価実験により,出力した語が病気の兆候・症状かを評価し,重み付けを行い,データベースを構築するという手法を提案する.構築したデータベースを用いることで,病気を予測するためのシステムの構築が望める.この提案手法を数式化すると以下のようになる.

①ユーザUから得られた 1 ツイートを,u_iとし入力とする.
②ツイートu_iに事実性解析を行い,正解ツイートをu_{it}とし,以下の処理を行う.
③正解ツイートu_{it}を投稿したユーザUの過去ツイートを,u'_{it}とする.
④過去ツイートu'_{it}に形態素解析を用いて,病気の兆候・症状として語$x(u'_{it})$を出力として取得する.
⑤評価実験として,出力$x(u'_{it})$と,後述する評価用のデータベースの語 1 語ずつを,コサイン類似度を用いてベクトル計算を行い,その最大値をc_iとする.
⑥最終的にc_iを出力$x(u'_{it})$の重みとし,以下のようなデータベースを構築する.

$$\begin{cases} x(u'_{1t}) : c_1 \\ \quad \vdots \\ x(u'_{it}) : c_i \end{cases} \cdots (1)$$

3.1 ツイートの事実性解析

事実性解析のプロセスでは,2 つの手法を用いて解析を行い,精度の高い結果を用いてデータの収集を行う.この手法の 1 つとして,SVM で学習データを用いてツイートの真偽判断を行う.もう 1 つの手法として,SVM の真偽判断の処理後,真と判断されたツイートに Zunda 解析器を用いたモダリティによる解析を行う.この 2 つの解析結果より,精度が高い結果を用いてデータの収集を行う.

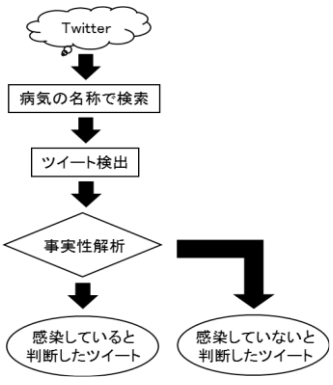

図 1:事実性解析のフローチャート

3.2 解析データを用いたデータ収集

①事実性解析の結果を用いて,データの収集を行う.事実性解析で”真”と判断されたツイートを投稿したユーザのそれ以前のツイートを対象に検出を行う.
②そのツイートの文章より,形態素解析を用いて病気の兆候・症状として,名詞・形容詞・動詞といった語を検出する.その語を,予め作成しておいた,病気に対しての兆候・症状をまとめた評価用のデータベースと比較し評価を行う.
③最終的に,取得した語に重みづけを行い,データベースの構築を行う.

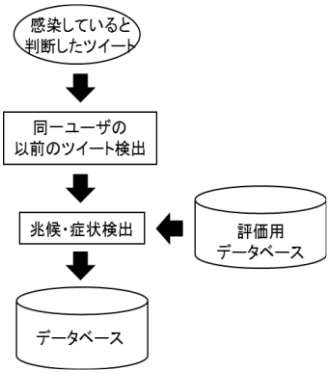

図 2:データ収集のフローチャート

3.3 アルゴリズムの詳細

提案手法の詳細を具体的な例を用いて示す.

(1) ツイート取得

はじめに,Twitter よりツイートを取得する.ここでは,ツイートを取得するために TwitterAPI を使用する.検索対象として,現在流行している病気の中から 10 種類の病気の名称,もしくはその同義語を用いる.同義語を検索対象として用いる病気は,病気の名称より同義語の方がツイート文中で使用されている傾向がある.

ここでは例として,「感染性胃腸炎」というキーワードを用いて検索を行い,実験方法を説明する.

図 3 は取得したツイートであり,これを病気毎に 300 ツイート取得し,事実性解析を行う.

感染性胃腸炎とやらで死にかけています
感染性胃腸炎だそうです。風邪じゃ無かった
@xxxx 感染性胃腸炎だよ、きっと…
大阪府枚方市 / ノロウイルスなどによる感染性胃腸炎にご注意
・
・
・

図 3: 感染性胃腸炎を対象に取得したツイート例

(2) 事実性解析

ここでは,取得したツイートに対して事実性解析を行う.事実性解析の手法として,SVM のみを用いた手法と,SVM の処理後 Zunda 解析器を用いたモダリティによる解析を行う.
①はじめに,SVM を用いた解析では,学習データと取得したツイートを用いて行う.ここで使用する学習データは,予め対象となる病気の名称で検索を行った 300 ツイートとする.このツイートを病気に感染しているか,感染していないかの真偽判断(ラベル付け)

を手動で行う.続けて,ツイート毎に形態素解析を行い,名詞・形容詞・動詞といった語を集合としてまとめる.
この学習データを用いて,取得したツイートに真偽判断(ラベル付け)を行う.取得したツイートは,学習データと同様に形態素解析を行い,名詞・形容詞・動詞といった語の集合としてまとめる.
図 4 は,学習データを用いて,取得ツイートに対してラベル付けを行う例である.この語集合と学習データを,SVM を用いて分類を行う.

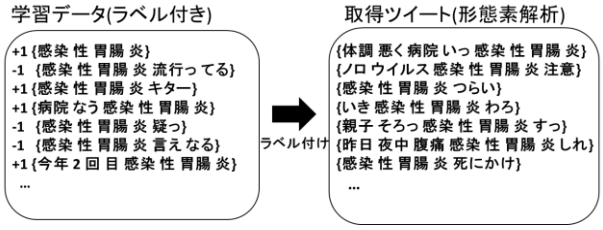

図 4:事実性解析例

②2 つ目の手法として,前述した SVM の解析結果より,真偽判断により”真”と判断されたツイートに対して,Zunda 解析器を用いる.Zunda 解析器を用いた解析の条件として,文章中で対象病気の名称が出現した箇所より,後の文章中で 1 番はじめに出現したイベントの真偽判断の結果を解析結果とする. 解析例として.「感染性胃腸炎でした。」という文章を解析した結果,モダリティにより,”真”となる.
③最終的に,2 つの手法で分類したそれぞれの結果の精度の高い方の結果を用いて,解析データを用いたデータ収集を行う.

(3) 解析データを用いたデータ収集

事実性解析によって,“真”と判断されたツイートより,それ以前に投稿された過去ツイートを検出する.過去ツイートを検出する際の条件として,取得ツイートより当日～5 日前の日付毎の範囲かつ,過去 100 ツイートの範囲でそれぞれ過去ツイートを検出する.過去 100 ツイートの範囲で検出する理由は,Twitter ではタイムライン上のツイートの話題が頻繁に変わるため,取得ツイートとの話題のつながりを保つために直近のツイートを取得する.
過去ツイートを検出後,形態素解析を行い,名詞・形容詞・動詞といった語を抽出する.その語を当日～5 日前の日付毎の範囲でまとめ,それぞれ病気の兆候・症状として扱う.

(4) 評価実験

過去ツイートに形態素解析を行い,取得した語に対して評価実験を行う.評価実験の手法の例を,図 5 に示す.評価用のデータベースと取得した語に,コサイン類似度を用いた類似度計算を行い評価する.
評価用のデータベースの作成方法として,病気に関しての症状・兆候等の情報がまとめられたサイト[4],[5]を用いる.サイト上の兆候・症状に関する文章に,形態素解析を用いて名詞・形容詞・動詞といった語を抽出し,それを評価用のデータベースとして扱う.
①取得した語 1 語に対して,評価用のデータベース 1 語ずつを,コサイン類似度を用いて類似度を算出する. コサイン類似度の算出は,式(2)を用いる.単語の文脈ベクトルは,単語中の 1 文字ずつの出現頻度を用いて算出を行う.取得した語に対するコサイン類似度で,最大の値をその語の評価値かつ重みとして扱う.

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} {q_i}^2}\cdot\sqrt{\sum_{i=1}^{|V|} {d_i}^2}} \cdots (2)$$

②取得した語に対しての評価値を用いて,類似した語の数,類似語の出現確率,類似度の平均を日付毎に出力する.
③最終的な結果として,取得した語とその重みをデータベースとして作成する.

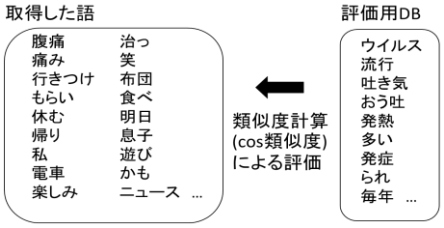

図 5:評価実験例

4. 実験

実験結果として,以下に対象とした 10 種類の病気に対して,事実性解析の精度,取得した語に対する評価をまとめた結果を示す.実験対象とした 10 種類の病気及び同義語を以下に示す.同義語を検索対象とした病気は()に名称を示す.
①感染性胃腸炎
②RS ウイルス(RS ウイルス感染症)
③アレルギー性鼻炎
④マイコプラズマ肺炎
⑤気管支炎
⑥水疱瘡(水痘)
⑦りんご病(伝染性紅斑)
⑧結膜炎(流行性角結膜炎)
⑨おたふく風邪(流行性耳下腺炎)
⑩喘息

4.1 事実性解析

実験を行った各病気に対して,それぞれ SVM を用いた解析と,SVM+Zunda 解析器を用いた解析結果を表 1 に示す.値の算出は式(3)で行う.SVM を用いた解析では,Classias[6]を用いて行う.
事実性解析の評価精度=(正解ツイート数)/(取得ツイート数)…(3)

表 1:事実性解析の評価精度

	SVM	SVM+Zunda解析器
感染性胃腸炎	57.77	56.33
RSウイルス	81.69	79.58
アレルギー性鼻炎	71.00	67.67
マイコプラズマ肺炎	75.33	74.33
気管支炎	61.00	58.33
水疱瘡	74.00	73.33
りんご病	65.33	72.33
結膜炎	67.00	68.33
おたふく風邪	78.67	76.67
喘息	65.00	66.33
平均	69.68	69.32

表の結果の平均より,SVMの解析がSVM+Zunda解析器の解析より,事実性解析の精度は高いと言える.かつ,りんご病等,一部の病気の場合は,SVM+Zunda 解析器の方が,精度が高いという結果が見られた.

4.2 実験結果

最終的に評価用のデータベースを用いて行った評価実験の結果をグラフに示す.値はそれぞれ,
①「抽出した語の数」は,過去ツイートから病気の兆候・症状として取得した語の数
②「類似した語の数」は,コサイン類似度を用いた類似度計算で閾値以上の値が出た語の数
③「類似語の出現確率」は,抽出した語に対しての類似した語の割合
④「平均類似度」は,コサイン類似度の平均値となっている.類似語の出現確率は式(4),平均類似度は式(5)によって算出する.

類似語の定義は,病気の兆候・症状と思われる語とする. 本実験ではコサイン類似度を用いた類似度計算により,算出した値が0.6より上の場合,その語は病気の兆候・症状と思われる語とする.
図 6 は,感染性胃腸炎の評価結果である.

類似語の出現確率[%]
= (類似した語の数)/(抽出した語の数) … (4)
平均類似度[%]
= (コサイン類似度の合計)/(抽出した語の数) … (5)

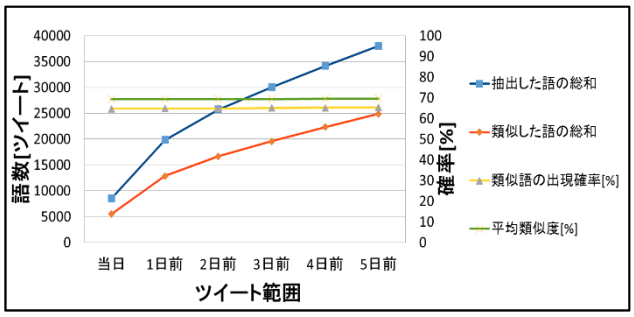

図 6:感染性胃腸炎の評価結果

図 6 より,類似語の出現確率及び,平均類似度は,日付の範囲によってはあまり変化しないことがわかった.この結果で得られた結果の精度を,重みとしてデータベースを作成する.表 2 は,作成したデータベースの一部である.

表 2:感染性胃腸炎の DB の一部

取得語	重み	出現回数	品詞
胃腸	1	43	名詞
感染	1	41	名詞
炎	1	41	名詞
性	1	41	名詞
風邪	1	29	名詞
ノロ	1	25	名詞

表 3 は,10 病気の評価結果を平均した結果である.表より,結果を平均した平均類似度は約 70[%]であった.

表 3:評価結果の平均

	抽出した語の総和	類似した語の総和	類似語の出現確率[%]	平均類似度[%]
平均値	19175.4	12733.5	66.30	69.30

4.3 考察

事実性解析の評価結果より,実験の対象とした 10 種類の病気で,SVM を用いた解析の方が,精度が高い結果は 7 種類,SVM+Zunda 解析器を用いた解析では 3 種類となった.この結果と評価の平均より,SVM を用いた解析の方が,SVM+Zunda 解析器より精度が高いと言える.この原因として,ツイート文章に対して Zunda 解析器を用いる際の条件が,より適切なものが存在するためだと考えられる.本実験で使用した条件の問題点として,文章中で 1 番はじめに出現したイベントの解析結果だけでは正確な真偽判断が行えてないことが考えられる.

SVM を用いた解析では,より精度を向上させるには学習データの増加,学習データ及び取得したツイートに対してのノイズ除去の条件を増加することが効果的だと考えられる.

解析データを用いたデータの収集の考察としては,病気の兆候・症状として取得した語より,算出した評価結果は日付毎にはあまり変化しないことがわかった.これにより,病気を発症する当日～5 日前の範囲では,病気の前触れとなるツイートが同じ頻度で投稿されているといえる.この実験を行う日付の範囲を,より拡大することで得られる病気の兆候・症状は,減少すると考えられる.これを行うことで,病気に感染していない時のツイート情報を取得できるため,最終目的としている病気を予測するための知識ベースの構築に,効果的だと考えられる.

5. 結論

結論として,実験により病気の兆候・症状としての語を取得することができたが,その中には病気の兆候・症状として相応しくない語が含まれていた.このことから,事実性解析及び,その結果を用いたデータ収集は,より精度の向上が望めると考えられる.

最終的に本研究の結果を用いて,病気を予測するためのシステムの作成を行うことが今後の課題となる.

謝辞

本研究は,JSPS 科研費 15K00425,15K00309,15K16077 の助成を受けたものです.

参考文献

[1][北川 2015]: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,言語処理学会,2015.

[2] Zunda:Japanese Extended Modality Analyzer
https://code.google.com/p/zunda/

[3][元田 2006]:元田浩,津本周作,山口高平,沼尾正行:データマイニングの基礎,pp.96-109. オーム社,2006.

[4] メディカル i タウン
http://medical.itp.ne.jp/

[5] 東京都感染症情報センター
http://idsc.tokyo-eiken.go.jp/

[6]Classias
http://www.chokkan.org/software/classias/index.html.ja