疾病サーベイランスのための罹患ツイートの自動獲得と事実性判定への利用

浅川 玲音
豊橋技術科学大学情報・知能工学課程
asakawa@nlp.cs.tut.ac.jp

秋葉 友良
豊橋技術科学大学情報・知能工学系
akiba@cs.tut.ac.jp

1 はじめに

マイクロブログは一般の多くの人々が発信する大量の情報を保有し，リアルタイム性に優れ，SNS としての特徴も持つとして近年注目されている情報源である．特に Twitter はマイクロブログの代表として様々な分析技術が研究開発されている [1]．そのような中で本論文では，Twitter を用いた疾患サーベイランスの研究領域に取り組む．

荒牧ら [2] は前述した特性に着目し，Twitter を用いたインフルエンザ流行予測手法を提案した．これは，まず Twitter から「インフルエンザ」に関連した発言を抽出し，次に SVM を用いた分類器でその発言者または発言者の付近の人物が実際にインフルエンザに罹患しているかどうかを判定するものである．この手法の中心である疾患の事実性判定技術は幾つか研究されているが，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．しかし人手によるラベリングはコストが高く，Twitter がツイートの配布を禁止しているのでラベリング済みツイートの共有も困難であるという問題点があり，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．

本研究では，簡単な方法で機械的に学習コーパスを収集し，tweet の投稿者またはその周囲の人間が何らかの疾患・症状に罹患しているか否かを判定する分類器を作成した．

2 提案手法

提案手法は，後述する罹患ツイートの自動獲得手法，学習コーパスの作成方法，ツイートの病気事実生判定のための分類器作成の３つに分けて説明する．

ツイートの分類ラベリングは NTCIR MedWeb task[3] の定義に従った．すなわち，あるツイートがその投稿者または周囲の人間が今現在何らかの疾患・症状に罹患していることを意味している場合，そのラベルを罹患 Positive であるとし，逆にそれを意味していない場合を罹患 Negative であるとした．表 1 に罹患 Positive の例と罹患 Negative の例を示す．

表 1: 罹患 Positive と罹患 Negative の具体例

咳つらいくるしいたすけてえ	Positive
頭痛が痛い昼寝しよ	Positive
頭痛が痛い（笑）よく言っちゃう www	Negative
春。俺、花粉症が酷いんだ.....	Negative

2.1 罹患ツイートの自動獲得

我々は Twitter の SNS 的な特徴，つまりユーザ同士の対話があることに着目し,「お大事に」というキーワードを使って『罹患ツイート (Symptom tweet)』の自動獲得を試みた [4].『罹患ツイート』とは「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のことを指す．人間同士の会話において「お大事に」と言われる相手は何らかの疾患・症状に罹患している．これを前提にし,『罹患ツイート』は罹患 Positive なツイートなのではないかという仮設を立て，Twitter API で収集した．収集期間は 2017 年の 5 月 1 日〜5 月 11 日，8 月 19 日〜9 月 3 日とし，合計 180,107 ツイートを入手した．

収集した罹患ツイートについて分析を行った．まず罹患ツイートを 50 件無作為抽出して人手で評価した結果，70.0%が罹患 positive であった．次にフィッシャーの正確検定を用いて罹患ツイートと後述する一般ツイートを比較し，罹患ツイート中の各単語が特徴的な使われ方をしているかどうかの分析を試みた．検定結果の p 値が小さい順に単語を取り出すと，図 1 に示すように罹患を表現するツイートに使われていそうな語

咳, 入院, 風邪, ひい, 病院, 体調, 頭痛, 喉, 薬, 痛い, 昨日, 今日, まし, た, GW, 痛く, 寝, 痛み, インフル, ので, て, 体調不良, 飲ん, 鼻水, 大事, 行っ, 早退, 心配, 大丈夫

図 1: 罹患ツイートに特徴的に現れる単語リスト

図 2: 罹患ツイートの特徴を表しているとされるカオモジや絵文字

が抽出されていることが確認できた．p 値が 0.005 未満（99.5%の確率で一般ツイートと異なる使われ方をしている）のものの中には，名詞以外の単語も多く含まれており，病気に関連性のある名詞だけでなく罹患の表現が広く取得できていることが考えられる．興味深いことに，図 2 のような特定のカオモジや絵文字も特徴的な使われ方をしていることも分かった．

2.2 学習コーパスの作成

罹患ツイートと併せて，検索キーワードを使わずにリンクがついていない日本語のツイートを一定期間収集した．これを体調不良を表していない，通常状態のツイートであると仮定し，『一般ツイート』と呼ぶことにする．収集期間は 2017 年 5 月 16 日〜2017 年 5 月 18 日とし，329,610 ツイートを入手した．

学習コーパスを作成するにあたり，まず罹患ツイートは全て罹患 Positive であり一般ツイートは全て罹患 Negative であると仮定することを考えた．しかしながら，罹患ツイートには 3 割の罹患 Negative なツイートが含まれており，一般ツイートは条件なく収集したものなので罹患 Positive を含んでいる可能性もあることが危惧される．そこで，罹患/一般ツイートのフィルタリングを試みた．

2.3 学習コーパスのフィルタリング

フィルタリングのために，既存の罹患ラベル付きコーパスである NTCIR-13 MedWeb タスク [3] で配布されたデータを利用した．このデータは，クラウドソーシングにより擬似的に作成したツイートから成り，予め定義した 8 疾患それぞれついて人手で罹患 Positive または Negative を判定したマルチクラスのラベルが

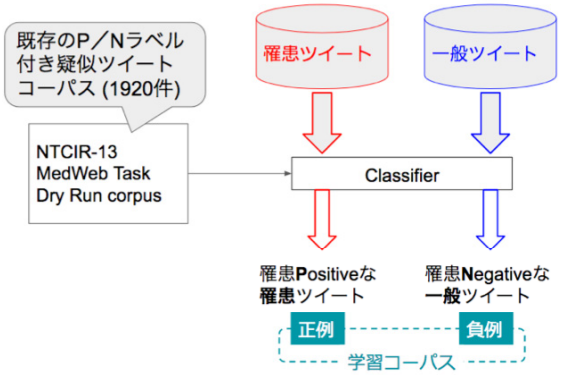

図 3: フィルタリングを施した罹患/一般ツイートからなる学習コーパス

「私は日本人です」という文の単語列

[私, 名詞, 代名詞, 一般,*,*,*, 私, ワタシ, ワ タシ], [は, 助詞, 係助詞,*,*,*,*, は, ハ, ワ], [日本人, 名詞, 一般,*,*,*,*, 日本人, ニッポンジン, ニッポンジン], [です, 助動詞,*,*,*, 特殊・デス, 基本形, です, デス, デス], [. , 記 号, 句点,*,*,*,*,. ,. ,.]

図 4: 分類器で使用する単語素性の例

付けられている．これを，一つ以上の疾患に罹患していれば罹患 Positive，そうでなければ罹患 Negative として，二値ラベルを付け直し，フィルター用の分類器の学習データとした．分類器は，MedWeb タスクのトレーニング用に配布された 1,920 件のツイートを学習データとして使用し，2.4 節で述べるツイートの病気事実性判定と同じ方法で構築した．

構築した分類器を用いて，学習コーパスのフィルタリングを行った (図 3)．収集した罹患ツイートのうち分類器で Positive であると判定されたツイートを正例，一般ツイートのうち分類器で Negative と判定されたツイートを負例として，新たな学習コーパスを作成した．正例については，分類器として使用した SVM での分類結果として，分離平面からの距離が一定以上離れているもののみを正例とするように厳しい条件を与えた．

2.4 ツイートの病気事実性判定

ツイートの分類器として，RBF カーネル，パラメータ γ=0.1，C=10 の SVM[5](Support Vector Machine) を使用した．素性はツイートを 2 値の Bag of

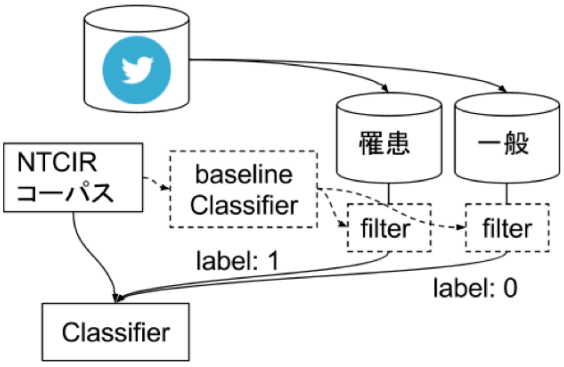

図 5: 分類器の学習

Words のベクトルで表現したものを用いた．ツイートは MeCab によって単語列に分割した．このとき，単語は MeCab(Neologd 辞書) による形態素解析結果に含まれる表層形・品詞情報・原形等の全ての情報を使用した．図 4 に例を示す．従って，基本的にはあるツイートにどの単語が含まれているかという情報のみが学習に利用されるが，MeCab による構造情報が少しだけ含まれている．語彙は罹患ツイート 180,107 件，一般ツイート 329,610 件，MedWeb タスクで配布されたトレーニング用コーパス 1,920 件に含まれる全ての語彙を使用した．語彙サイズは 234,105 語となった．また学習データの不均衡への対策として SMOTE[6] を使用した．

3 病気事実性判定評価実験

本実験では，収集したツイートの「学習データとしての質」「既存のコーパスのデータを拡張する能力」「フィルタリングによる質の向上」の 3 点を評価する．具体的には，作成したフィルタリングあり/なしコーパスのみを学習した分類器の性能と，既存コーパスに作成したフィルタリングあり/なしコーパスを併せて学習した分類器の性能を，ベースライン分類器の性能と比較した．

3.1 評価方法

評価のために，NTCIR-13 MedWeb task[3] のラベル付けガイドライン [1] に従って，実際のツイートに罹患 Positive か Negative かを人手で判定したテストデータを作成した [2]．具体的には，NTCIR と同じ 8 疾患名をキーワードにして各疾患について約 80 件ずつ実際のツイートを 639 件収集し，人手で罹患 Positive か Negative かを二値判定し，ラベリングを施した．収集したツイート 639 件のうち，357 件が罹患 Positive，282 件が罹患 Negative であった．

評価指標には，罹患 pos/neg の予測結果と正解ラベルを比較して得られた，F 値と正解率 (Accuracy) を使用した．

3.2 比較手法

学習データの違いによる分類性能を比較した．比較した学習データを以下に示す．これらを単独，あるいは組み合わせて，分類器の学習データを構築した．分類器は 2.4 節で述べた方法で構築し，全て共通である．

NT 既存の罹患ラベル付きコーパスである MedWeb のトレーニング用ツイートコーパス 1,920 件 (正例 1,390 件，負例 530 件)．

RE 提案手法で収集した罹患ツイートを正例，一般ツイートを負例として構築したコーパス．データサイズは，既存コーパスとあわせて，1,920 件 (正例 1,390 件，負例 530 件) とした．

preRE 提案手法で収集したツイートをフィルタリングして作成したコーパス．罹患ツイートのうち罹患 Positive と判定されたものを正例，一般ツイートのうち罹患 Negative と判定されたものを負例とした．データサイズは，既存コーパスとあわせて，1,920 件 (正例 1,390 件，負例 530 件) とした．

RE*2 RE のサイズを 2 倍にしたコーパス 3,840 件 (正例 2,780 件，負例 1,060 件)．

preRE*2 preRE のサイズを 2 倍にしたコーパス 3,840 件 (正例 2,780 件，負例 1,060 件)．

提案手法と比較するベースラインは，既存のコーパスである **NT** のみを学習データに使用した分類器である．これと，提案手法で収集したデータを単独で用いた場合 (**RE**, **preRE**)，および既存コーパスと組み合わせて学習データを拡張して使用した場合 (**NT+RE**, **NT+RE*2**, **NT+preRE**, **NT+preRE*2**) の比較を行った．

[1] http://mednlp.jp/medweb/NTCIR-13/doc/ja-ver2.0.pdf

[2] MedWeb タスクのテストデータに対しても評価を行ったが，提案法の効果は確認できなかった．MedWeb タスクは擬似的に作成したツイートを対象としているため，収集した実際のツイートとのミスマッチがあるのではないかと考え，本実験では実際のツイートを対象とするように設定を行った．

表 2: 事実性判定評価実験結果

学習データ	**NT**	**RE**	**preRE**	**NT+RE**	**NT+RE*2**	**NT+preRE**	**NT + preRE*2**
F 値	**0.727**	0.708	0.724	0.725	0.728	0.732	**0.743**
Accuracy	0.581	0.573	**0.610**	0.587	0.595	0.609	**0.632**

3.3 実験結果

実験結果を表 2 に示す.

まず，同じサイズのコーパスを学習データに用いた **NT**, **RE**, **preRE** を比較する．**NT** と **RE** の比較により，自動獲得した学習データ (**RE**) は，既存のラベル付きデータ (**NT**) の質には及ばないことが分かる．しかし，**RE** と **preRE** の比較から，フィルタリングによりコーパスの質を改善できることが分かる．また，**NT** と **preRE** の比較から，フィルタリングにより自動獲得したデータでも既存のラベル付きデータと同程度の分類性能が得られることが分かる．

次に既存のラベル付きデータに自動獲得したデータを追加することによる効果を調べる．**NT** と **NT+**α の比較から，提案手法によって収集したコーパスで既存コーパスのデータ拡張を行うことで分類器の性能が向上することが分かる．また，**NT+RE** と **NT+preRE**，および **NT+RE*2** と **NT+preRE*2** の比較から，フィルタリングを行うことでデータ拡張の効果をさらに向上させることができることが分かる．さらに，**NT+RE** と **NT+RE*2**，および **NT+preRE** と **NT+preRE*2** の比較から，学習データを増やすことで分類性能をさらに改善できることが分かる．

4 おわりに

本研究では,「Tweet の投稿者またはその周囲の人間が何らかの疾患に今現在罹患しているか否か」の分類に取り組んだ．具体的には，罹患を決定づける表現が様々であるのに対して学習コーパスを十分に用意することは困難であるという問題の解決策として，罹患ツイートの自動獲得手法を提案し，獲得したツイートを用いて機械的に作成したコーパスで学習した分類器による病気事実性判定性能を評価した．

評価結果として，作成したコーパスを既存コーパスのデータ拡張に利用することで分類性能を向上させることができること，フィルタリングによって作成したコーパスの質を向上させることができることを示した．従って自動獲得した罹患ツイートが病気事実性判定に利用可能であると結論付けられる．今後の研究では，フィルタの改良，素性や分類器の変更などによる病気事実性判定能力のさらなる向上を目指す．

また，提案手法を日本語以外のツイートに適用することも検討したい．我々の予備実験では，日本語の「お大事に」に相当する英語の “get well soon” でも英語の罹患ツイートを収集できるが,「お大事に」に比べると収集した罹患ツイートの精度が低いことを確認している．対象の言語に適したキーワードを選択したり，複数のキーワードを組み合わせて精度・再現率を上げる工夫が必要であると考えている．

参考文献

[1] 奥村学. ソーシャルメディアを対象としたテキストマイニング. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, Vol. 6, No. 4, pp. 285–293, 2013.

[2] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1576, 2011.

[3] Eiji Aramaki, Shoko Wakamiya, and Mizuki Morita. Overview of the NTCIR-13: MedWeb task. *Proceeding of the NTCIR-13 Conference*, pp. 40–49, 2017.

[4] Asakawa Reine and Akiba Tomoyoshi. AKBL at the NTCIR-13 MedWeb task. *Proceedings of the NTCIR-13 Conference (2017)*, pp. 52–55, 2017.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, Vol. 20, No. 3, pp. 273–297, September 1995.

[6] N. V. Chawla, L. O.Hall K. W. Bowyer, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, pp. 321–357, 2002.