ウェブからの疾病情報の大規模かつ即時的な抽出手法

荒牧英治 * **
篠原（山田）恵美子 ****
森田瑞樹 ***
岡瑞起 *

* 東京大学 知の構造化センター
** 科学技術振興機構 さきがけ
*** 独立行政法人 医薬基盤研究所
**** 東京大学 医学部附属病院

eiji.aramaki@gmail.com
emiko-tky@umin.ac.jp
mizuki@bi.a.u-tokyo.ac.jp
mizuki.oka@gmail.com

1　はじめに

近年，フェイスブック[1]やツィッター[2]などのマイクロブログにより，多くの人々の情報が大量に利用可能である．これにともない，地震や台風などこれを利用した研究も盛んに行われている．我々は疾患情報に注目し，ツィッターから全国的な疾病状態の収集を目指している．マイクロブログを利用した疾病状態の収集には次の 2 つの利点がある

【大規模】疾病情報は通常，医療施設からの報告を収集して行われる．例えば，厚労省のインフルエンザ流行レベルマップ[3]は全国 5,000 の医療機関の定点観測の集計である．一方，ツィッターでは毎日数万を超えるインフルエンザに関する書き込みが投稿されており，大規模な情報収集が可能となる．

【即時性】先のインフルエンザ流行レベルマップは 1 週間間隔の更新であり，非常事態においてその察知が遅れる可能性がある．一方，ツィッターにおいては，任意の時間での情報を集計可能であり，超早期での警告が可能である．

以上のような利点はあるもののマイクロブログを情報源とするには，様々なノイズから正しい情報を選別する必要がある．例えば，単に「風邪」という表現が含まれている発言を収集すると以下のような発言が収集される：

(1) もし，彼が**風邪**だったら中止しよう
(2) **風邪**なんでしょうか？
(3) **風邪**の特効薬ができればノーベル賞もんだ

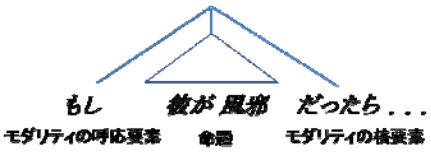

図 1: モダリティによって事実性を持たない文.

これらはそれぞれ以下に上げる理由で「風邪をひいている」という事実をもたない．(1)は単に風邪の可能性を仮定しているだけである（図 1）．(2)は風邪について質問しているだけである（疑問文）．(3)は風邪について言及されているものの研究対象としての風邪であり，そこには<*風邪をひいた*>という事実はない．

このように疾患情報の抽出を行うためには，単語<風邪>を含んだ発言の中から＜本人が風邪をひいた＞発言のみを抽出する必要がある．本稿では，この<本人が疾病状態にある>ことを疾病状態の**事実性**と呼ぶ．

事実性がない場合の原因は様々であるが，本研究では，モダリティが適切でない場合とそもそも対称となる命題が言及されていない場合の大きく 2 種類があると考える．例えば，先の例では(1)と(2)は可能性や疑問文といったモダリティが原因で非事実となっているとみなす．一方，(3)には，そもそも疾患としての風邪がないため，命題自体が不適切であり，(1)(2)とは異なった現象である．このように，両者は性質が異なった現象であり，これらを区別して扱った方が，自然な学習が可能である可能性がある．また，将来的にはデータの可搬性も高まると考える．

本研究のポイントは 2 つある．

(1) 疾患情報の収集のために，事実性を判定するタスクを提案する．

[1] http://www.facebook.com/
[2] http://twitter.com/
[3] https://hasseidoko.mhlw.go.jp/Hasseidoko/Levelmap/flu/

(2) モダリティと命題を別個に識別し，これを組み合わせることで，事実性を判定する手法を提案する．

2 事実性コーパス

まず，作成したコーパスについて述べる．扱うタスクは<風邪>とその諸症状，<喉の痛み>など，計 7 つのタスクである(表 1)．それぞれのタスクについて次のようにコーパスを構築した．

【STEP1：データ収集】 表 1 の検索クエリをもとに Twitter クロールデータ[1]を検索し，TASK1 は 5000 件，TASK2-7 は 1000 件を収集した．

【STEP2：事実性アノテート】発言毎にタスクが事実(+1)または 非事実(-1)のフラグの付与を行った．ここでいう事実とは「本人または本人のまわり（同じ都道府県）でその疾患を持っている人が現在存在する，また，近い過去存在していた」か否かで判断を行った．事実とみなす基準の例を表 2 に載せる．

【STEP3：非事実原因アノテート】STEP2 にて非事実と判定された発言の原因がモダリティによるものか(モダリティ=+1)，そうでないか(モダリティ=-1)を判定し，フラグを付与した．ここでいうモダリティを表 3 に示す． モダリティが原因でなく，そもそも命題が不適切な場合のフラグ付与(命題=-1)も同時に行った．構築したコーパスの例を表 4 に示す．

3 提案手法

本手法の基本アイデアは，命題部とモダリティ部で分けて学習を行う点である．まず，基本となる発言の識別手法について述べてのち(3.1 節)，これを用いた提案手法を述べる(3.2 節)

3.1 基本となる文識別手法

基本となる発言の識別手法は，検索クエリ周辺の右コンテキスト（後方）の形態素と左コンテキスト（前方）の形態素 n-gram を bag-of-words として扱い，正解ラベルを学習する．

正解ラベルは，提案システムでは命題とモダリティと 2 つあるため，2 つの識別器を用いる．事実性を直接学習するベースラインのシステムは 1 つの識別器のみを用いる．学習は Support vector machine (SVM) [2]を用いる(図 2)．ウィンドウ幅は図 3 の予備実験の結果左右 6 形態素とした．カーネルは 2 次多項カーネルを用いた[4]．

[4] TinySVM 実装による；実行オプション svm_learn -t 1 -d 2 -c 1

ID	タスク	検索クエリ
TASK1	風	風邪
TASK2	喉の痛み	喉の痛み，のどの痛み
TASK3	寒気	寒気，悪寒，さむけ
TASK4	鼻水／鼻づまり	鼻水，鼻づまり，鼻風邪
TASK5	咳／たん	咳，痰
TASK6	熱	高熱，微熱，発熱
TASK7	頭痛	頭痛

表 1: タスクとコーパス収集に用いた検索クエリ．

	状況	発現例
+1	治っている最中の場合	***だいぶ風邪も落ち着いた***
−1	治ってから 1 日以上の経過した場合	***先週の風邪はひどかった***
+1	身近な人や直接会った人に症状が出ている	***社内で風邪が流行ってるみたい***
−1	一般論	***風邪なみなさんは、耳鼻科へ行くといいです***

表 2: 事実性アノテーション基準.
* 詳細な基準は http://mednlp.jp/~aramaki/KAZEMIRU/ を参照されたい．

モダリティ	分類	発現例
法	**仮定法**	***風邪だったとしても行く***
時制	**過去時制**	***昨年はひどい風邪で参加できませんでした***
表現類型	**疑問文**	***風邪でしょうか？***
	命令文	**風邪ひいちゃえ**
価値判断	**必要許可**	**風邪でもいい**
仮想	**可能世界**	***風邪で行けないという夢をみた***

表 3: 事実性を損なうモダリティ．

タスク	A	B		検索クエリ	発言
	事実性	モダリティ	命題		
風邪	-1	-1	+1	風邪	たとえ風邪でも休めない
	+1	+1	+1	風邪	風邪で頭痛い
	-1	+1	-1	かぜ	かぜで病院が激混み
熱	-1	-1	-1	微熱	微熱だったとしても行く！
	-1	-1	-1	発熱	発熱でしょうか？

表 4: タスクとコーパス収集に用いた検索クエリ．
*STEP2 にて A 部を，STEP3 にて B 部のフラグを作成する．

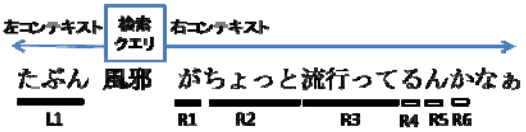

図 2: SVM の素性となる形態素列.

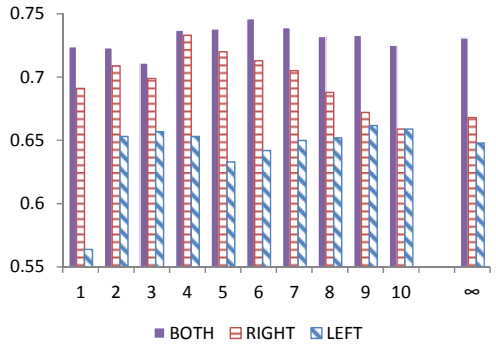

図 3：ウィンドウサイズと精度（F 値）．
RIGHT は命題の後方，LEFT は命題の前方，BOTH は両方．数字は形態素数を示す．BOTH∞は文のすべての形態素を用いる．BOTH6(前方後方の両方の 6 形態素を用いる)の精度が最も高いことから，以降の実験ではこれを用いた．

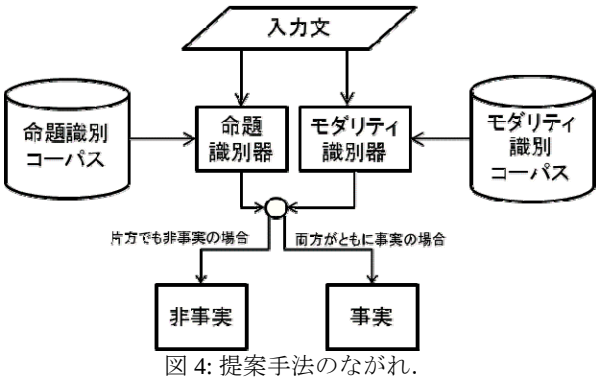

図 4: 提案手法のながれ．

	手法 A	手法 B
TASK1	0.825 (p=0.781,r= 0.781)	0.828 (p= 0.811,r= 0.838)
TASK2	0.962 (p=0.933,r=0.962)	0.962 (p= 0.934,r=0.993)
TASK3	0.661 (p=0.675,r=0.648)	0.670 (p=0.681,r=0.659)
TASK4	0.834 (p=0.839,r=0.829)	**0.847** (p=0.836,r=0.859)
TASK5	0.869 (p=0.829,r=0.914)	**0.879** (p=0.816,r=0.953)
TASK6	0.689 (p=0.698,r=0.679)	**0.705** (p=0.665,r=0.750)
TASK7	0.906 (p=0.877,R=0.936)	**0.912** (p=0.866,r=0.964)

表 5: 実験 1 の結果（精度は F 値による）．
括弧内 p は適合率，r は再現率を示す．太字は手法 A と手法 B で有意に精度向上した箇所を示す（マクネマー検定; p=0.05）．

	手法 B	+ TASK1 ALL	+ TASK1 モダリティ	+ TASK1 命題
TASK2	0.962	0.954	0.954	0.947
TASK3	0.670	0.678	**0.713**	0.647
TASK4	0.847	0.871	**0.892**	0.830
TASK5	0.879	0.861	0.880	0.880
TASK6	0.705	**0.759**	0.711	0.739
TASK7	0.912	0.910	0.907	0.914

表 6: TASK1 データの転移結果．
太字は手法 B と比較し有意に精度向上した箇所を示す（マクネマー検定; p=0.05）．

3.2 識別器の組み合わせによる手法

提案手法の枠組みを図 4 に示す．本研究では，事実性が成り立つためには，命題部とモダリティ部の両方が適切であると仮定している．そこで，モダリティを識別する識別器（モダリティ識別器）と命題の識別器（命題識別器）を別個に構築し，両方が事実と判定した場合を事実とし，一方でも非事実の場合には非事実とする．

この手法は，単に事実性を学習するのに比べ，コーパス構築の手間がかかるという欠点があるが，本来別個の現象であるならば，別個に学習することで，より自然な学習が可能だと期待される．また，命題部やモダリティ部の性質が近いタスク間ではドメインアダプテーションの際の有効なリソースとなりうる．

4 実験

命題部とモダリティ部を組み合わせて事実性を判定できるかどうかを調査した (実験 1)．また，モダリティ部分と命題部分はタスク固有なデータであるのか，それとも他のタスクでも利用可能であるのかを調査した (実験 2)．

4.1 実験 1:モダリティ部と命題部を分けて学習可能か？

提案手法（モダリティ部と命題部の 2 つ事実性判定の組み合わせ）と従来の事実性判定の精度を比較した．これは次の 2 つの手法の精度を比較することで行った．

【比較手法】事実性を直接学習する手法 A（ベースライン）と提案手法の手法 B を比較した．
【評価】適合率，再現率，F 値(β=1)による．
【結果】結果を表 5 に示す． F 値では手法 B が 7 つのタスク中 4 つで有意に高く，事実性を直接学習するよりも，モダリティと命題と分けて学習を行った方が最終的なパフォーマンスがよくなると言える．ただし，提案手法の学習データ作成の方が時間とコストがかかるため，向上した精度とのバランスが考慮されるべきである．

また，命題とモダリティがそれぞれ別個に扱った方がパフォーマンスがよいということは，これらは機械学習的にもクリアに切り分け可能な現象だとも考えられ，本研究の仮定を支持する．

4.2 実験 2:データは可搬性を持つか？

あるタスク，例えば風邪，について事実性判定のコーパスを構築行っても，他のタスクを行う際には新たなコーパスを再度構築する必要があるならば多くの時間とコストがかかる． そこで構築したデータが他の疾患や症状を推定するタスクに転用可能かどうかを調査した．これは，TASK1(風邪)のタスクにおいて，他のタスクのコーパスと合わせて学習した場合，精度がどう

変化するかを調べることで行った．本実験ではコーパスがタスクに依存するかどうかを調査するのが目的であるため，ドメインアダプテーションで一般的な手法（データの重みの調整やデータ由来の付加[3]）を用いず，単にコーパスを結合する手法を用いた．

【比較手法】 4つの手法を比較した．

(1) **手法 B**: 前節のシステム．
(2) **+TASK1ALL**: (1)に TASK1 のデータを加えたシステム．
(3) **+TASK1 モダリティ**: (1)に TASK 1 のモダリティデータのみを加えたシステム．
(4) **+TASK1 命題**: (1)に TASK1 の命題データのみを加えたシステム．

【他の設定】 学習器，評価法，素性は実験 1 と同じとした．

【結果】 結果を表 6 に示す． 2つのタスクにてモダリティ部分のデータの可搬性の効果があることが示された．しかし，他のデータの可搬性が効果を示す場合やそもそも効果がない場合もある．以上のことから，疾患に関する表現であっても，他の疾患には容易に転移できないことが分かる．しかし，2 つのタスクで，モダリティ部分の転移が成功しているため，モダリティ部分はタスク間で共通要素が多い可能性がある．今後の考察を深めたい．

5 関連研究

モダリティの一部，特に否定(Negation)や疑い(Suspicion)は，専門のワークショップ[4]が開催されるなど，盛んに研究されてきた．特に，医療分野では古典的な研究トピックであり，今世紀の早い段階から研究がある．Chapman 等[5]は否定を捉える正規表現によるアルゴリズムを提案した．Goldin 等[6]は機械学習（ナイーブベイズと決定木）により Negation を学習した．Elkin 等[7]は否定のスコープの先頭と末尾の語のリストを記述した．Veronika 等[8]は negation のスコープを記述したコーパスを構築した．これらの研究では，述べられた命題に対しての相対的な事実性を扱っている．すなわち，「風邪が治った」といった場合，「風邪が治る」ことは事実であるので，事実と判定する立場をとっている．一方，本研究で扱う事実性では，疾患に対しての事実性であり，研究の観点が異なる．また，日本語においては，江口等[9]が態度表明者，時制，仮想，態度， 真偽判断，価値判断，焦点などについて詳細に事象アノテーションを行っている．本研究は事実か非事実のみの観点から上記をまとめたものとみなせる．

6 まとめ

本研究ではツィッターから全国的な疾病状態の収集を行った．これを実現するためには，発言から疾病があったという事実の有無を判定する手法が必要となる．提案手法では事実でない場合を(1)命題が不適切である場合と(2)モダリティが不適切である場合に分けて識別し，一部のタスクにて精度を高めた．ただし，コーパス作成にコストがかかるため，今後，データの可搬性を高め再利用可能にすることが課題である．

謝辞:本研究は，JST 戦略的創造研究推進事業（さきがけタイプ）「情報環境と人」及び，科研費補助金(若手研究 A)による．材料の整備にあたっては，エスエス製薬(株)， UTIX(株), 及び McCANN HEALTHCARE WORLDWIDE JAPAN の協力による．

参考文献

[1] 荒牧英治, 増川佐知子: 微小時間における日本語の変化とその法則, 言語処理学会 第 17 回年次大会, 2011.

[2] Vladimir Vapnik,: The Nature of Statistical Learning Theory, Springer-Verlag, 1999.

[3] Hal Daume, III: Frustratingly Easy Domain Adaptation, In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 256-263.

[4] NeSp-NLP 2010 : Negation and Speculation in Natural Language Processing.

[5] Wendy Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce Buchanan. 2001b. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 5:301-310.

[6] Ilya M. Goldin and Wendy Chapman. 2003. Learning to detect negation with not in medical texts. In Workshop at the 26th ACM SIGIR Conference. Yang Huang and Henry J. Lowe. 2007. A novel hybrid approach to automated negation detection in clinical radiology reports. Journal of the American Medical Informatics Association, 14(3):304-311.

[7] Peter L. Elkin, Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom, and Dietlind L. Wahner Roedler. A controlled trial of automated classification of negation from clinical notes. BMC Medical Informatics and Decision Making 5:1.

[8] Veronika Vincze, Gyorgy Szarvas, Richard Farkas, Gyorgy Mora, and Janos Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics, 9(11).

[9] 江口萌, 松吉俊, 佐尾ちとせ, 乾健太郎, 松本裕治: 日本語文章の事象に対する判断情報アノテーション, 情報処理学会研究報告, 自然言語処理研究会, 2009-NL-193, 2009.