風邪に罹ったのは誰か？ — 疾患・症状を保有する主体の推定

叶内 晨 * [1]　小町 守 [1]　岡崎直観 [2,4]　荒牧英治 [3,4]　石川 博 [1]

[1] 首都大学東京　[2] 東北大学　[3] 京都大学　[4] 科学技術振興機構さきがけ

1 はじめに

ソーシャルメディアの普及により，個人の意見・体験が言語ビッグデータとして蓄積されるようになった．同時に，言語ビッグデータを通して個人や社会の意見・状況を集約しようとする試みも進められている．荒牧ら [5, 6] は，「風邪」や「インフルエンザ」などの疾患・症状を含むツイートから，個人やその周辺人物が疾患・症状に罹っている発言を選別し，風邪やインフルエンザの流行状況を把握するシステムを提案した．

荒牧らのシステムの解析誤りの分析を進めたところ，「誰が疾患・症状にあるのか」という，疾患を保有する主体の推定が重要であることが判った．例えば，「娘が風邪を引いた」という発言において「風邪」という疾患を保有するのは「娘」であることが解析できれば，発言者の近くで「風邪」が出現したことが分かる．一方，「風邪と風を誤変換していた」という発言では「風邪」という疾患を保有している主体が存在せず，風邪の流行とは無関係である．つまり個人の状態を把握するためには，調べたい状態に言及する表現を認識することに加え，その状態に置かれている人物の特定が重要である．

自然言語処理においてこのタスクに最も近いのは，述語項構造解析である．もし，調べたい疾患・症状が事態性名詞である場合（例えば「発熱」）は，そのガ格を調べればよい．しかしながら，疾患・症状が事態性名詞になるかどうかは，述語項構造解析のアノテーション基準に依る所が大きく，通常「風邪」「鼻水」などは事態性名詞として扱われない．

代わりに，用言の項構造に着目するアプローチも考えられる．先ほどの「娘が風邪を引いた」という例では，「風邪」は「引いた」のヲ格で，「娘」は「引いた」のガ格なので，「風邪」の保有者は「娘」と推定できる．しかし，このアプローチにも複数の問題がある．第１に，風邪を保有していることを表す述語を識別する問題である．例えば「医者が風邪を診察した」という文では，「風邪」は「診察した」のヲ格で，「医者」は「診察した」のガ格であるが，「風邪」の保有者は「医者」ではない．第２に，口語表現特有の解析誤りがある．例えば「風邪引いた」のようにヲ格が省略されると，述語項構造解析は失敗してしまう．さらに，「風邪ツラい」などカタカナの表現は，形態素解析にすら失敗する可能性がある．このように，既存の述語項構造解析の研究と，疾患・症状を保有する主体を推定するタスクの間には，かなりの乖離がある．

本論文では，疾患・症状を保有する主体を推定するという新しいタスクに取り組む．まず，ツイートの本文に対して，疾患・症状を保有する主体をラベル付けしたコーパスを構築するための方針を設計し，アノテーション作業を行った．構築したデータを訓練事例として用い，疾患・症状を保有する主体を推定する解析器を設計した．評価実験では，主体を推定する解析器の精度を計測すると共に，主体を推定することによる後続のタスク（疾患の流行を認識するタスク）での貢献を実証した．また，疾患・症状の主体を推定するタスクは，個別の疾患・症状への依存することなく，一般的な解析器を構築できることが分かった．

2 コーパス

本研究では，荒牧ら [5] によって作成されたコーパスを利用する．このコーパスは，「風邪・咳・頭痛・寒気・鼻水・熱・喉の痛み」の７種類の症状に関して，ツイートの発言者が疾患・症状にあるかどうかをラベル付けしたものである [1]．各症状を含む発言は，ツイート検索で収集されており，例えば「熱」に関する発言は「発熱」「微熱」「高熱」をクエリとして収集された．

このコーパスでは，投稿者が発言した 24 時間前までに疾患・症状にあったと判断できる発言に正例，それ以外の発言に負例ラベルが付与されている．ただし，症状を保有する人物が話者以外の場合でも，その人物が話者と同じ都道府県にいると判断できる場合には，症状は正例と判定されている．これはカゼミル＋ [2] において，都道府県別に疾患の流行状況の集約をしているためである．コーパスサイズは，風邪コーパスのみ 5,000 tweet で，他の疾患コーパスはそれぞれ 1,000 tweet 前後である．

*kanouchi-shin@ed.tmu.ac.jp

[1]「喉の痛み」のみ，エラー分析を行った結果，負例が一定数に達しなかったために実験の対象から外した．

[2] http://kazemiru.jp

表 1: 疾患症状に関する主体ラベルの種類と発言例

ラベル	意味	発言例
一人称	発言した話者が疾患・症状に関与	風邪引いてひきこもりたい
周辺人物	話者が直接見聞きできる範囲の人物が疾患・症状に関与	弟がめっちゃ咳してて怖い
その他人物	それ以外の人物が疾患・症状に関与	@*****　風邪ですか？ お大事に。。。
物体	人間以外の物体や生物が状態の主体	また PC が発熱
主体なし	主体が存在せず，疾患のイベントが発生していない	だるいし、風邪薬買って帰る〜

表 2: 疾患クエリを保有する tweet の主体ラベルの比率

ラベル名	一人称	周辺人物	その他人物	物体	主体なし	合計
tweet 数	2153	129	201	40	401	2924
tweet 内に主体あり	70	112	175	38	0	395
正例：負例	1833：320	99：30	2：199	0：40	16：385	1950：974

図 1: ラベル付けの例

本研究では，荒牧ら [5] のコーパスに誰が疾患・症状にあるのかの情報を付与した．この作業は，各疾患毎に 500 件ずつ行った．ラベルの種類と発言例を表 1 に，ラベル付けの例を図 1 に示す．ソーシャルメディアの分析では，一次情報（本人が観測・体験した情報）であるかどうかの識別が重要なので,「一人称」「周辺人物」「その他人物」「物体」「主体なし」の 5 つのラベルを用意した．

「一人称」のラベルの，発言した話者が疾患・症状に関与する という意味は，必ずしも症状にある場合だけではなく，主体が症状の保有に関係する場合を全て含む．例えば，表 1 の一人称の発言例のように症状に対して願望を抱いている場合は，今は症状を保有していないため，カゼミル＋の応用から考えると抽出したくない情報である．しかし，本研究は疾患・症状を保有する主体を推定することに重きを置いているので,「一人称」のラベルを付与する．主体が「周辺人物」「その他人物」「物体」の場合にも同様な条件で判断し，主体ラベルを付与した．

「周辺人物」のラベルは話者が直接見聞きできる範囲の人物が症状にあるかを一つの分類基準とした．荒牧らの疾患コーパスは，話者か話者と同じ都道府県の人物が症状にある場合に正例となるが，人手で主体のラベル付けする際に，同じ都道府県かどうかを判断することは極めて難しいためである．

「その他人物」のラベルは，症状を保有する主体となる人物が存在するが,「一人称」「周辺人物」「物体」には該当しない全てのケースを含む．返信先に症状の主体が存在する場合が一例で，表 1 の発言例では話者と物理的に見聞きできる距離にいることを確認できない．

「物体」のラベルは物体，もしくは人間以外の生物が主体となる場合に付与され，パソコンなどの物体が発熱した場合が例として挙げられる．

「主体なし」のラベルは，発言例にある「風邪薬」のように，風邪が名詞句の一部として出現する場合を含む．他にも「寒気」が「さむけ」ではなくて「かんき」として使われるような語義が異なる場合や，疾患・症状が慣用句的に使われている場合，記事・作品のタイトルとして出現する場合にも「主体なし」とした．

また，主体推定の難しさを分析するため，発言中に主体を示す表現がある場合は，その表現にもラベルを付与した．表 2 を見ると，主体が「一人称」の場合にはほぼ省略されるが，主体が「周辺人物」「その他人物」「物体」の場合には約 9 割が発言内で言及される．荒牧らの正解ラベルとの比率に着目すると,「一人称」と「周辺人物」の場合は約 8 割が正例である一方で,「物体」「その他人物」「主体なし」の場合は 1 割以下であった．

3 実験

3.1 主体推定器

2 節のコーパスを利用して，発言内での「風邪」や「頭痛」などの疾患・症状を保有している主体を推定する分類器を構築する．今回の実験では,「物体」と「主体なし」のラベルを「主体なし」に統合した．なお，1 つの発言に疾患・症状が複数言及されている場合と，同じ疾患・症状を保有する主体が複数存在する場合は，学習事例から取り除いた．ツイート中のリツイート，返信，URL は，有無のフラグを残した上で削除した．分類器には Classias 1.1[3] を利用し，L2 正則化ロジスティック回帰モデルを学習した．利用した素性を表 3 に示す．

3.2 推定結果

表 4 に，5 分割交差検定により主体推定の精度を測定した結果を示す．訓練事例として，6 つの疾患・症状に関するコーパスをマージした 3,000 事例を用いた．全ての素性を組み合わせた結果，macro F1 スコアはベースライン（BoW）と比べて約 20 ポイント上昇した．これ

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

表 3: 主体推定器の素性

Bag-of-Words (BoW)： 疾患クエリの前後 9 個の内容語の表層形.
疾患クエリ (Query)：疾患クエリが何か.（例：風邪）
2,3gram：疾患クエリの前後 6 文字を 2 文字，3 文字ずつ連結させた文字 2gram，文字 3gram.
URL： 発言内に URL があるかどうか.
RP, RT：その発言に返信（リプライ）・リツイート・非公式リツイートがあるかどうか.
周辺人物辞書 (Ndict)：周辺人物の主体として適切な単語を人手で集め [4]，それらが発言内に一つでもある場合に発火（例：彼女・社員・部下）
その他人物辞書 (Odict)：上記の Ndict と同様にして，その他人物辞書を作成し使用.（例：幼児）
人名 (OnesName)：「さん・君・ちゃん」の正規表現と一致したものと，mecab の解析結果で人名が発言内にある場合に発火.
TweetSize：発言の形態素の長さ毎に，10 個以下，11 個から 30 個，31 個以上の 3 つに分けた素性.
疾患クエリが主辞 (IsHead)：疾患クエリの次の形態素が名詞以外の時に疾患クエリが主辞であるとして発火 [5].

は，提案した素性がうまく作用していることを示唆している．疾患クエリ，リプライの有無，周辺人物辞書，が特に強い貢献を示した．

表 5 に予測と正解の Confusion Matrix を示す．対角成分の太字の数値は予測が正解したケースである．(+数字) はベースラインと比べ，予測した事例数が何件変化したかを表す．例えば「周辺人物」の推定は 34 件成功し，ベースラインからは 22 件増加している．

表 4 において macro F1 スコアが micro F1 スコアより低い理由として，主体ラベルの正解比率の問題が挙げられる．表 5 の右端の列より，「一人称」の主体が全体の約 7 割を占めることが分かる．この比率により，分類器のバイアス項の重みは「一人称」に傾き，主体推定器は「一人称」のラベルを付与しやすくなっている．よって，「一人称」のラベルの再現率が高い一方で，その他のラベルの再現率は低下している．その結果，発言数の少ない「周辺人物」「その他人物」「主体なし」の推定性能が伸び悩み，macro F1 スコアが低下している．

3.3 主体推定における疾患・症状への依存性

3.2 節の実験では，6 つの疾患・症状に関する全ての訓練事例を利用した．では，疾患・症状を保有している主体を推定するタスクは，どのくらい個別の疾患・症状に依存するのか？　本節では，風邪に関する訓練事例のみを用いた場合と，すべての疾患・症状に関する訓練事例を用いた場合の性能を比較する．

図 2 は風邪の主体を推定する際に 5 分割交差検定を行った結果を示している．実線は風邪コーパスのみを用

[4] ラベル付けしたコーパス全体を見た上で作成した．その他人物辞書においても同様にした．

[5] 疾患クエリの次の名詞が「気味」など特定の場合は例外とした．

表 4: 主体推定の素性と精度

素性	micro F1	macro F1
BoW (baseline)	0.772	0.422
BoW+ Query	0.819	0.536
BoW+ 2,3-gram	0.791	0.461
BoW+ URL	0.773	0.427
BoW+ RP,RT	0.800	0.471
BoW+ Ndict	0.776	0.468
BoW+ Odict	0.773	0.427
BoW+ OnesName	0.771	0.427
BoW+ TweetSize	0.774	0.433
BoW+ IsHead	0.776	0.435
全ての素性	**0.840**	**0.618**

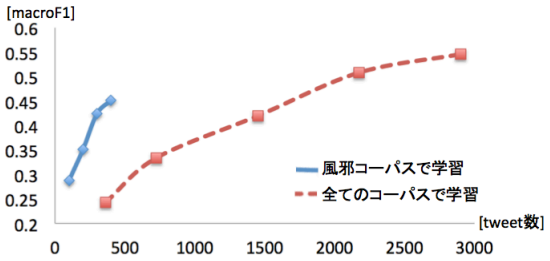

図 2: コーパスサイズと推定精度

いて学習した場合，点線は全てのコーパスで学習した場合の性能である．全てのコーパスの学習を行う際には，風邪コーパスを 100，200，300，400 件と増やすと同時に，風邪以外のコーパスもランダムに 625，1,250，1,875，2,500 件増やしている．風邪の主体を予測するタスクなので，風邪に関する学習データとの相性がよく，400 件の学習データを用いた場合の F1 スコアは 0.451 であった．一方，風邪以外の症状に関する学習データを追加し，2,900 件の訓練事例を用いて風邪の主体を予測した場合の F1 スコアは 0.546 で，風邪のみの学習データを用いた場合と比較すると 9.5 ポイント向上した．

風邪の主体を予測するだけであれば，風邪に関する訓練事例を増やすことが最も効果的であるが，風邪以外の疾患・症状の主体に関する訓練事例を増やすことで，特定の疾患・症状だけに依存しない汎用的な主体推定器を構築できる可能性が示唆された．同様の傾向は，他の疾患・症状を予測対象とした場合でも確認された．

ただ，疾患・症状を保有する主体の事前分布にばらつきがあるため，疾患・症状の依存性が皆無という訳ではない．例えば，頭痛に関する言及では 9 割以上の主体が一人称の頭痛のことを表すが，熱に関しては物体の状況（例えば PC の発熱など）を言及するものも多い．したがって，幅広い疾患・症状をカバーしたコーパスを構築し，主体推定器の汎用性を改善していく必要がある．

3.4 疾患・症状判別器の結果

本研究の大元の目的である，疾患・症状の流行を認識するタスクにおいて，本研究で構築した主体推定器がどのくらい貢献するのか，実験を行った．表 6 は本論文で

表 5: 主体ラベルの予測と正解の Confusion Matrix

予測ラベル			一人称		周辺人物		その他人物		主体なし		正解の合計
正解ラベル	一人称	i	**2,089**	**(−10)**	6	(+1)	20	(+16)	38	(−7)	2,153
	周辺人物	n	85	(−15)	**34**	**(+22)**	5	(−4)	5	(−3)	129
	その他人物	f	96	(−41)	6	(+0)	**81**	**(+38)**	18	(−3)	201
	主体なし	o	184	(−148)	2	(+1)	9	(+3)	**246**	**(+144)**	441
予測の合計			2,454	(−14)	48	(+24)	115	(+53)	307	(+137)	2,924

表 6: 疾患・症状判別器の素性と F 値

	風邪	咳	頭痛	寒気	鼻水	熱	macro F1
ベースライン（BL）[F]	0.844	0.885	0.908	0.759	0.892	0.781	0.845
BL+推定した主体 [F]	0.850	0.883	0.907	0.814	0.894	0.802	0.858
BL+ゴールドデータの主体 [F]	0.877	0.926	0.935	0.885	0.914	0.886	0.904

提案した主体推定器を利用して主体ラベルを推定し，その主体ラベルを素性に追加して症状の有無を判定した結果である．なお，ベースライン手法は荒牧ら [5] の素性設計を参考にして，本研究で独自に実装し，それぞれの症状ごとに 500 tweet を 5 分割交差検定した．推定した主体を素性として利用した結果，寒気の F1 スコアが 5.5 ポイント，熱の F1 スコアが 2 ポイント向上し，全体の macro F1 スコアも 1.3 ポイント向上した．

本研究で付与した主体の正解ラベル（ゴールドデータ）を素性として利用した場合とベースラインを比較すると，「風邪・咳・頭痛・鼻水」は F1 スコアで 2〜4 ポイント程度向上し，「寒気・熱」は 10 ポイント以上向上した．これにより主体を正しく判定することができれば，平均で約 6 ポイントの F1 スコアの向上が見込める．本研究で構築した主体推定器により，特に「寒気・熱」において，ゴールドデータとの差を縮めることができた．寒気の精度が向上した理由のひとつには，「寒気」が「さむけ」ではなく「かんき」として使われる場合や，「悪寒」が「予感」として使われる場合を排除できたことが挙げられる．

4 おわりに

本論文では，複数の疾患・症状に関して，その症状を保有する主体を推定する取り組みを紹介した．構築したコーパスを訓練事例とした主体推定器を作成し，主体の推定が micro F1 スコアで 84 ポイント程度の性能で行えること，異なる疾患・症状に対して横断的な主体の推定が可能であることを報告した．さらに推定した主体が疾患を判定する上でどの程度貢献するのか実証した．

関連研究としては，動詞の名詞化に着目し，PropBank [2] に準拠した項構造を事態性名詞に付与した NomBank [1] がある．日本語では，京都大学テキストコーパス 4.0[6] や NAIST テキストコーパス [7] において，事態性名詞の項が付与されている．小町ら [7] は，名詞に事態性があるか否かの事態性判別と，その後の項同定を別タスクとして扱い，解析精度を報告している．また，「娘の風邪」などの名詞句内の関係を解析する研究 [4] も関連がある．発言内で疾患・症状の主体が省略されていることも多いため，省略・照応解析 [3] とも関連がある．

しかしながら，疾患・症状を保有する主体を推定するというタスク設定に完全に一致する関連研究はない．ソーシャルメディア上の発言から個人の状態を分析することは，疾患の流行予測・把握のみならず，個人の健康状態をモニタリングするなどの重要な応用がたくさんある．今後は，さらに多くの疾患・症状でも検証を進め，ソーシャルメディア上の投稿に対する自然言語処理の主要なタスクとして育てていきたい．

[6] http://nlp.ist.i.kyoto-u.ac.jp
[7] https://sites.google.com/site/naisttextcorpus/

参考文献

[1] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank Project: An interim report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, pp. 24–31, 2004.

[2] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106, 2005.

[3] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *COLING*, pp. 769–776, 2008.

[4] Ryohei Sasano and Sadao Kurohashi. A probabilistic model for associative anaphora resolution. In *EMNLP*, Vol. 3, pp. 1455–1464, 2009.

[5] 荒牧英治, 森田瑞樹, 篠原恵美子, 岡瑞起. ウェブからの疾病情報の大規模かつ即時的な抽出手法. 言語処理学会第 17 回年次大会発表論文集, pp. 838–841, 2011.

[6] 荒牧英治, 増川佐知子, 森田瑞樹. 文章分類と疾患モデルの融合によるソーシャルメディアからの感染症把握. 自然言語処理, Vol. 19, No. 5, pp. 419–435, 2012.

[7] 小町守, 飯田龍, 乾健太郎, 松本裕治. 名詞句の語彙統語パターンを用いた事態性名詞の項構造解析. 自然言語処理, Vol. 17, No. 1, pp. 141–159, 2011.