WebNLP: NLP 応用の誤り解析

～事実性解析と主体解析を題材に～

荒牧英治[1,2] 叶内晨[3] 北川善彬[3] 岡崎直観[4,2]

1) 京都大学 2) 科学技術振興機構 3) 首都大学東京 4) 東北大学

1. はじめに

今世紀に入って以降，自然言語処理（NLP）は Web とともに発達してきた．Web を大規模な構造化テキストコーパスと見なし，そこから知識や統計量を抽出することで，形態素解析，構文解析，固有表現抽出，述語項構造解析，機械翻訳など，様々なタスクで精度の向上が報告されている．これらの事例は，Web が NLP を高度化したと捉えることができる．

同時に， Web は誰もが発信できるメディアであり，その特性を活かした Web ならではの新しい研究分野も形成された．評判情報抽出(Pang, Lee et al. 2002)がその代表例である．さらに，近年では，Twitter や Facebook などのソーシャルメディアが爆発的に普及し，さらに Web に関する関心が高まっている．

ソーシャルメディアのデータには，（1）大規模，（2）即時性，（3）個人の経験や主観に基づく情報など，というこれまでに無い特徴がある．例えば，「*熱が出たので病院で検査をしてもらったらインフルエンザ A 型だった*」という投稿から，この投稿時点（即時性）で発言者は「インフルエンザに罹った」という個人の経験を抽出し，大規模な投稿の中からこのような情報を集約できれば，インフルエンザの流行状況を調べることができる．このように，NLP で Web 上の情報をセンシングするという研究は，地震検知(Sakaki, Okazaki et al. 2010)，疾病サーベイランス(Parker, Wei et al. 2013)を初めとして，選挙結果予測，株価予測など応用領域が広がっている．

ただし，これらの応用事例は実用性を強く意識しているため，NLP として似たような課題を複数の応用事例で個別に解いてしまい，NLP の研究としての積み重ねが行われていない，という問題点が生じている．そこで，我々 WebNLP プロジェクトでは，次のゴールを設定した．

（1） ソーシャルメディア上のテキストの蓄積を自然言語処理の方法論で分析し，人々の行動，意見，感情，状況を把握しようとするとき，現状の自然言語処理技術が抱えている問題を認識すること

（2） 応用事例（例えば疾患状況把握）の誤り事例の分析から，自然言語処理で解くべき一般的な（複数の応用事例にまたがって適用できる）課題を整理すること

（3） （2）で見出した個別の課題に対して，最先端の自然言語処理技術を適用し，新しいタスクに取り組むことで，自然言語処理のソーシャルメディア応用に関する基盤技術を発展させること

表 1： 検索のためのキーワード

	症状	キーワード
風邪コーパス	風邪	*風邪*
	寒気	*寒気，悪寒，さむけ*
	鼻水	*鼻づまり，鼻水，鼻ずまり，鼻風邪*
	咳	*咳，痰*
	熱	*高熱，微熱，発熱*
	頭痛	*頭痛*
インフルエンザ・コーパス	インフルエンザ	*インフル*

本プロジェクトでは，NLP によるソーシャルリスニングを実用化した事例の１つである，ツイートからインフルエンザや風邪などの疾患・症状を認識するタスクに取り組む．本報告書の第２節では，このタスクの説明を行い，第３節で既存のシステムの誤り分析を行う．分析結果から，事実性の解析，状態を保有する主体の判定が重要かつ一般的な課題として切り出せると判断し，第４節，第５節ではこれらの課題に取り組んだ成果を報告する．第６節で本プロジェクトの結論を述べる．

表 2：誤り分析（False Positive についての誤りの分類）

	風邪	咳	頭痛	寒気	鼻水	熱
FN:FP の比	30:70	25:75	15:85	58:42	25:75	36:64
FP の分類と頻度（頻度の多いものから）	非当事者（19）	非当事者（26）	比喩（20）	比喩（33）	比喩（22）	非当事者（31）
	時制（10）	話題（14）	話題（16）	非当事者（3）	話題（15）	時制（10）
	話題（10）	モダリティ（10）	モダリティ（15）	否定（1）	モダリティ（9）	話題（8）
	モダリティ（9）	時制（9）	非当事者（13）	モダリティ（1）	非当事者（8）	比喩（4）
	否定（7）	否定（8）	時制（7）	時制（1）	時制（6）	否定（3）
	比喩（6）	比喩（2）	否定（4）		否定（2）	モダリティ（2）
	その他（12）	その他（11）	その他（11）	その他（4）	その他（19）	その他（5）

2. コーパス

本研究では，風邪およびその症状に関する Twitter 上での発言を集めたコーパス（以下，**風邪症状コーパス**）と，インフルエンザに関する Twitter 上での発言を集めたコーパス（以下，**インフルエンザ・コーパス**）の２つを用いる．風邪症状コーパスは，誤り分析及び，主体解析の検証のため，インフルエンザ・コーパスは，事実性判定の検証のため用いる．

先行研究においても，風邪やインフルエンザなど感染症に関する研究は多く(Parker, Wei et al. 2013)，他にも西ナイル熱 (Sugumaran and Voss 2012)などが扱われている．これらの多くは，経験則により「*風邪*」や「*インフルエンザ*」などのキーワードとなる単語を選択し，その頻度を集計し，感染状況の把握を行っている(Culotta 2010; Aramaki et al. 2011)．本研究では，先行研究の１つである(Aramaki, Maskawa et al. 2011)で使われたコーパス，及び，商用サイト「カゼミル・プラス」[1]で用いられたコーパスを用いる．

これらは 2008 年 11 月から 2010 年 7 月にかけて Twitter API を用いて 30 億発言 を収集し，次に，そこから「*インフルエンザ*」や「*風邪*」といったキーワードを含む発言を抽出したものである（表 1）．

2.1 風邪症状コーパス

風邪症状コーパスは，「*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（**陽性**か**陰性**か）をラベル付けしたものである[2]．

このコーパスでは，投稿者が以下の除外基準に照らし，1 つでも該当するものがあれば陰性とみなした．

- 発言者（または，発言者と同一都道府県近郊の人間）の疾患でない．居住地が正確に分からない場合は陰性発言とみなす．例えば，「*風邪が実家で流行っている*」では，「*実家*」の所在が不正確であるので，陰性とみなす．
- 現在または近い過去の疾患のみ扱い，それ以外の発言は除外する．ここでいう「*近い過去*」とは 24 時間以内とする．例えば，「*昨年はひどいインフルエンザで参加できなかった*」は，陰性とみなす．
- 「*風邪でなかった*」等の否定の表現（ネゲーション）は陰性とする．また，疑問文や「*かもしれない*」といった不確定な発言も陰性とする．

コーパスサイズは，風邪コーパスのみ 5,000 発言で，他の疾患コーパスはそれぞれ 1,000 発言である．

2.2 インフルエンザ・コーパス

インフルエンザ・コーパスは，「インフルエンザ」を含む約 10,000 件の発言に対して，陽性か陰性かをラベル付けしたものである．アノテーションの基準については，風邪症状コーパスに準拠している．なお，正例数が 1,319，負例数が 9,124 となっている．

3. 誤り分析

風邪症状コーパスを用いて誤り分析を行った．誤りを検出するために，ベースラインとして，単語 n-gram 素性を用いた文献(Aramaki, Maskawa et al. 2011)と同等の分類器を SVM にて構築し，その誤りを人手で分類した．

誤りには，本来，疾患の事実があると判定すべきであるのに，それができなかった場合（False Negative）と，その逆に，疾患の事実がないのに誤って疾患とみなしてしまう場合（False Positive）がある．ここで，

[1] http://kazemiru.jp/

[2]「喉の痛み」は，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．

前者の False Negative（以降，FN）となった事例に関しては，機械学習の学習結果の内部が分からない限り正確な誤りの原因を推定することが困難である．しかし， False Positive（以降，FP）と判定された事例に対しては，なぜ，それが Negative なのかという観点から，比較的容易に誤りを考察可能である．よって，FP について，人手で誤りを分類した．これは，風邪症状コーパスの 6 つの症状について，それぞれ 100 件の誤り事例について行った（600 事例）．結果を表 2 に示す．誤りの分類は以下の通りである．

- **非当事者**：疾患をもつ対象が，発言者およびその周辺の人物でない．
- **時制**：疾患のあった時間が異なる．
- **話題**：そもそも疾患の話題ではない．
- **モダリティ**：「*かもしれない*」（疑い），「*かな？*」（疑問）などのモダリティにより疾患の事実が認められない．
- **否定**：「*風邪でなくてよかった*」など，疾患の事実が否定されている．
- **比喩**：比喩としての疾患表現．
- **その他** ：その他の理由による．

このように現象としては 7 つの誤りに分類可能である．ここで，これを言語処理の研究課題という観点からまとめなおすと，疾患があったのかという**事実性の問題**（時制，モダリティ，否定，比喩，話題）と，仮に疾患の事実があったとして，疾患をもった対象が誰なのかという**主体の問題**（非当事者の問題や話題の問題）という２つの大きな言語現象に大別できる．

ここで，**事実性の問題**は Web 文章のみならず，言語処理全般に共通した問題である．

また，**主体の問題**も，疾患に限ったことではなく，評判抽出（だれの評判なのか？），感性情報処理（だれが喜び／悲しんでいるのか？），など Web 文章，特にブログなど個人が発言する情報を扱う上で基盤となる技術であり，Web を扱うために解くべき大きな問題である．

本稿では，これら２つの言語処理の課題に対して，それぞれ新規に取り組み，改善した結果を報告する（4 章にて事実性解析，5 章にて主体解析）．

4. 事実性解析

4.1 事実性解析が必要な事例

インフルエンザの流行情報の検出のために，機械的に「インフルエンザ」を含む発言を集めるだけでなく，文に記述されている事象が，実際に起こったことなのか，そうでないことなのかの事実性を判定する技術が必要となる．これは，**事実性解析**と呼ばれる技術である．事実性解析が必要な例は以下のような例である．

(1) 熱があったので、病院に行ったらインフルエンザだった。
(2) インフルエンザに罹ったかもしれない。
(3) インフルエンザに罹っていたら、休まざるを得ないだろう。

上記例のうち，インフルエンザの事実があったのは（１）のみであり，（２）は「かもしれない」という推量，（３）は「たら」という仮定であり，インフルエンザが存在した事実を持たない．このような事実化どうかを決定している波線の表現はモダリティと呼ばれ，事実性判定においてで重要な手がかりになる．

このため，Web 応用のタスクでは，文献 (Li, Ritter et al. 2014)がモダリティを陽に扱った素性を利用している．また，特にモダリティの一部である否定 (Negation) や疑い (Suspicion) については，専門のワークショップ[3]が開催されるなど盛んに研究されてきた．

本研究では，インフルエンザ流行検出のために，モダリティに関する既知のリソースやツールをフルに利用して精度を検討する．インフルエンザ・コーパスを用いた実験の結果，事実性解析においてモダリティを利用することで 3.5 ポイントの精度向上がみられ，事実性解析におけるモダリティの重要性を示した．

4.2 手法

「インフルエンザ」を含むウィンドウを中心として，左側の 3 つの形態素と右側の 3 つの形態素を Bag of Words（BoW）の素性とし分類器をベースラインとした．素性は以下を扱った．

- **window6BoW**: 「インフルエンザ」を含むウィンドウを中心とする左側3つの形態素と右側 3 つの形態素 の Bag of Words 素性
- **URL**: 発言における URL の有無の素性
- **Atmark**: 会話等によるリプライの有無素性

[3] http://www.clips.ua.ac.be/NeSpNLP2010/

表 3: 事実性解析の実験結果.

素性の組み合わせ	適合率	再現率	F 値
window6BoW	0.740	0.305	0.432
window6BoW+URL	0.699	0.313	0.432
window6BoW+Atmark	0.740	0.305	0.432
window6BoW+N-gram	0.707	0.345	0.464
window6BoW+Season	0.724	0.333	0.456
window6BoW+tsutsuji	0.764	0.321	0.452
window6BoW+Zunda	0.699	0.313	0.432
baseline	0.697	0.392	0.502
baseline+tsutsuji	0.702	0.420	0.526
baseline+Zunda	0.679	0.412	0.513
All (baseline+tsutsuji+zunda)	0.689	0.440	0.537

表 4: 重みの絶対値の大きい意味 ID 素性とその表層系例

ID	重み	表層形の例	ID	重み	表層形の例
D11	0.55	ないと駄目	z25	0.88	ではどう
l31	0.54	あまりに	k51	0.69	には
r12	0.47	事にしてる	h11	0.68	について
s23	0.41	おかげで	R11	0.55	みたい
G21	0.37	ことにする	I23	0.46	だろう
i11	0.36	といった	A31	0.44	そう
I12	0.34	らしい	I11	0.43	かも

表 5: Zunda による重みの絶対値の大きい素性

正の素性	**重み**	**負の素性**	**重み**
罹患=成立	0.80	注射=成立	0.62
かかり=成立	0.65	対策=成立	0.50
診断=成	0.52	かかり=0	0.48
寝=成立	0.47	なる=成立	0.45
診断=成立	0.52	する=成立	0.45
発覚=成立	0.47	死亡=成立	0.42
回復=成立	0.44	行っ=成立	0.39
ダウン=成立	0.40	注意=成立	0.38
うつっ=成立	0.39	感染=不成立	0.37
潜伏=成立	0.37	なっ=不成立	0.34

- **N-gram**: インフルエンザ」の前後の文字 N-gram の素性.前後の文字 1-gram から 4-gram の素性
- **Season**: 12 月から 2 月にかけてのインフルエンザ流行中の発言なのかそうでないかの素性

「つつじ」による素性

モダリティに関するリソースとして「つつじ」（日本語機能表現辞書）[4]の利用を試みた．日本語の文を構成する要素には，主に内容的な意味を表す要素（内容語）以外に，助詞や助動詞といった，主に文の構成に関わる要素がある．ここでは，後者を総称して，「機能語」と呼び，「に対して」や「なければならない」のように，複数の語から構成され，かつ，全体として機能語のように働く表現である「複合辞」と合わせて

[4] http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/

これらを機能表現と呼ぶ．つつじは 16801 の機能表現の表層形を階層的に分類しており，同じ意味を持つ機能表現には同じ意味 ID が振られている．

なお，本タスクは Twitter のデータを使用しており，発言の中には文が複数ある場合がある．これにより，注目しているインフルエンザ感染に関連する機能表現と関係のない機能表現も多く存在すると考えられる．そこで，「インフルエンザ」の右の 15 文字中につつじの機能表現の表層形が含まれる場合にその意味 ID を素性として利用した．

「Zunda」による素性

モダリティに関するツールとして，「Zunda」（拡張モダリティ解析器）[5]を利用した． Zunda は文中のイベント（動詞や形容詞，事態性名詞など）に対して，その真偽判断（イベントが起こったかどうか），仮想性（仮定の話かどうか）などを解析することのできる日解析器である．本手法においては，Zunda の出力する真偽判断のタグを素性として利用した．真偽判断についてのラベルには，「成立」，「不成立」，「不成立から成立」，「成立から不成立」，「高確率」，「低確率」，「低確率から高確率」，「高確率から低確率」，「0」のラベルが存在する．

さらに，これらのラベルがインフルエンザに関連するかどうかを考えなければならない．我々は Zunda が動詞，事態性名詞を「イベント」として解析していることから，「インフルエンザ」の右に続く動詞，事態性名詞で一番近いものをインフルエンザに関連するイベントとみなし，そのイベントとラベルの組み合わせを素性として利用した．

4.3 実験

実験はインフルエンザ・コーパスを用い， 5 分割交差検定による適合率，再現率， F 値で行った．ツールとしては，Classias (ver.1.1)[6]を使用した．ウィンドウを決めるための形態素解析器としては MeCab (ver.0.996)[7]を利用し，辞書は IPA-Dic (ver.2.7.0) を用いた．

結果を表 3 に示す．まず，window6BoW にそれぞれの素性を 1 つだけ加えたものを表上部に示す．URL の素性とリプライによる素性 Atmark を加えたときは，window6BoW に比べて，変化はないが，その他の素性については全て F 値が向上した．特に，向上が見られたのは，文字 N-gram，Season に関しての素性であった．全体の傾向として素性を加えた場合，適合率はほとんど下がらないが，再現率が上がっていく傾向が見られた．本タスクにおいては負例の割合が非常に大きく，適合率を上げるのは難しいと考えられる．

「つつじ」による素性の効果

つつじの意味 ID を用いた素性を window6BoW に加えたところ，F 値が 2.2 ポイント向上した．ここで注目したいのは，この素性により，適合率，再現率が共に上がっている点である．また，モダリティ以外の素性を全て合わせた baseline の素性につつじの意味 ID を加えると 2.4 ポイントほど向上が見られた．この場合も，適合率，再現率が共に上昇している．

「Zunda」による素性の効果

Zunda による素性を window6BoW に加えたところ，僅かの向上しか認められなかった．しかし，モダリティ以外の素性を全て合わせた baseline の素性に Zunda による素性を加えたところ，F 値は 1.1 ポイントほど向上した．

なお，つつじと Zunda に関しての素性を両方用いた場合，最高精度となった．この結果は，モダリティを用いなかい baseline より，3.5 ポイントの F 値の向上があり，モダリティに関する素性が有用であることを示せた．

4.4 考察

事実性解析において，どのようにモダリティに関する素性が貢献したかを考察する．

つつじにおける素性について，分類器の判断に大きく影響を与える素性を調べた．その結果を表 4 に示す．推量における「らしい」などが負の重みになっており，学習が適切に行われていることが分かる．

しかし，一部の発言においては，ひらがな 1 文字のものが多くマッチしてしまい「と」や「え」などはつつじの意味 ID「Q11」，「I23」に該当し，分類に失敗することとなった．

次に，Zunda による素性について，分類器の判断に大きく影響を与える素性を調べた．つつじの場合と同様に，重みの大きな素性を大きい順に並べた結果を表 6 に示す．これらの多くは直感的に理解できるものが多い．インフルエンザの発言では，注意を呼びかける発言，予防接種の内容の発言，ニュースに関する発言等が多く，負の重みによくそれが現れている．正の重みに関しては直接疾患に関係のある名詞や動詞が多くなっている．

---

[5] https://code.google.com/p/zunda/

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

[7] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

表 6. 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7. 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

表8. 主体推定の素性と精度

素性	microF1	macroF1
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 + RT	0.780	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**

このように，モダリティの素性を陽に組み込むことで，精度があげることができることを実証的に示した．

5. 主体解析

5.1 主体解析が必要な事例

はじめに述べたように，Web データをフルに利用するためには，事実性解析とならんで，誰が疾患・症状にあるのかという主体の推定（**主体解析**）が重要である．

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となる．このように，Web の情報を利活用するためには，その状態を所有している人物の特定が重要となる．

従来の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，述語項構造解析が失敗してしまう．

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，乖離がある場合がある．

本節では，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．次に，この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する．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最終的なタスク（疾患の流行を認識するタスク）での貢献を実証した．

5.2 手法

風邪症状コーパスにおいて，誰が疾患・症状にあるのかの情報を付与した．この作業は，疾患毎に 500 件ずつ行った．ラベルの種類と発言例を表 6 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 6 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，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．
- **「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る．例えば，＜発熱＞では，発熱しているのが，人間でなく，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である．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．

5.3 実験１：主体解析

風邪症状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 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．利用した素性を以下に示す．

表9. 主体ラベルの予測と正解の Confusion Matrix

出力		一人称	周辺人物	その他人物	主体なし	合計
正解	一人称	**2089 (−10)**	**6** (+1)	20 (+16)	38 (−7)	2153
	周辺人物	85 (−15)	**34 (+22)**	**5** (−4)	5 (−3)	129
	その他人物	96 (−41)	6 (+0)	**81 (+38)**	**18** (+3)	201
	主体なし	184 (−148)	2 (+1)	9 (+3)	**246 (+144)**	441

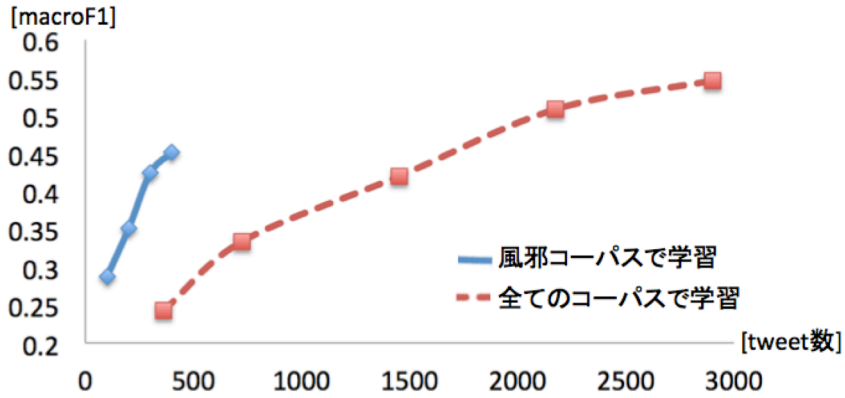

図1:コーパスサイズと推定精度

表10. 疾患・症状判別器の素性と F 値

	風邪	咳	頭痛	寒気	鼻水	熱	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]	0.85	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

- **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．(例：風邪)
- **2, 3gram**：疾患クエリの前後 6 文字を 2 文字，3 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，それらが発言内に一つでもある場合に発火させた．(例:彼女・社員・部下)
- **その他人物辞書（Odict）**：Ndict と同様にして，その他人物辞書を作成し使用した．(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.

表 8 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．特に，疾患クエリ，リプライの有無，周辺人物辞書が強い貢献を示した．

表 9 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．

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．例えば，「一人称」が全体の約 7 割を占めることから，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．

5.4 実験２：コーパスの可搬性

主体の判定は一般的なタスクであるならば，ある疾患に関する主体判定のデータを他のコーパスにも利用可能であるはずである．そこで，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．コーパス毎の相違点としては，例えば，「風邪」と「引く」の共起頻度は高いが，「頭痛」と「引く」の共起頻度は低い．よって，ドメインの異なるコーパスを利用した場合の精度向上は自明ではない．

図 1 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す．実線は風邪コーパスのみを用いて学習した場合，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であることから，当然，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．

4.4 実験３：風邪への貢献

最終的な目的である，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を調べる実験を行った．表 10 は本論文で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 3 章と同等である．学習事例は 6 つの症状においてそれぞれ 500 tweet ずつ利用し、5 分割交差検定を行った．

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．

6. おわりに

本稿では，NLP によるソーシャルリスニングを実用化した事例の 1 つである，ツイートからインフルエンザや風邪などの疾患・症状を認識するタスクに取り組んだ．分析結果から，事実性の解析，状態を保有する主体の判定が重要かつ一般的な課題として切り出せると判断し，これらの課題を陽に扱った手法により実験した結果，両問題がそれぞれ疾患・症状の認識に貢献することが明らか

となった．今後，両解析技術が発展し，より深く Web テキストを扱うことが期待される．

参考文献

Aramaki, E., S. Maskawa and M. Morita (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. EMNLP**:** 1568-1576.

Li, J., A. Ritter, C. Cardie and E. Hovy (2014). Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts. EMNLP**:** 1997-2007.

Pang, B., L. Lee and S. Vaithyanathan (2002). Thumbs Up? Sentiment Classification Using Machine Learning Techniques. EMNLP.

Parker, J., Y. Wei, A. Yates, O. Frieder and N. Goharian (2013). A framework for detecting public health trends with Twitter. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining**:** 556-563.

Sakaki, T., M. Okazaki and Y. Matsuo (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web**:** 851-860.

Sugumaran, R. and J. Voss (2012). Real-time spatio-temporal analysis of West Nile virus using Twitter data. Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications**:** 1-2.